Česká zemědělská univerzita v Praze

Provozně ekonomická fakulta

Katedra Informačních Technologií

Diplomová práce

Web 3.0

Bc. David Macháček

! ! !

Místo této strany vložíte zadání diplomové práce.
(Do jedné vazby originál a do druhé kopii)

! ! !

Čestné prohlášení

Prohlašuji, že svou diplomovou práci "Web 3.0" jsem vypracoval samostatně pod vedením vedoucího diplomové práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce. Jako autor uvedené diplomové práce dále prohlašuji, že jsem v souvislosti s jejím vytvořením neporušil autorská práva třetích osob.

V Praze dne 31.3.2013 _____

Poděkování

Rád bych touto cestou poděkoval ing. Alexandru Vasilenko za připomínky, ochotu a vstřícnost během tvorby této práce.

Web 3.0

Web 3.0

Souhrn

Cílem práce je nastínit význam pojmu web 3.0. Definovat jeho jednotlivé vlastnosti se zaměřením na nejdůležitější aspekt, tedy proces sémantizace. Práce se zabývá způsoby, jakými je tento proces realizován a jaké pro to používá technologie.

Tyto přístupy jsou následně realizovány v praxi a na základě výsledků dojde k vybrání optimálního nástroje pro implementaci strukturovaných dat v rámci webového dokumentu.

Práce se zabývá architekturou sémantického webu. Rozebírá jeho jednotlivé složky a způsoby, kterým jsou propojeny do jednoho fungujícího modelu, který umožňuje následné využívání strukturovaných dat v nich obsažených.

Je třeba identifikovat výhody i hrozby, které sémantický web přinese. Nalézt možnosti jeho využití i obory, kterým je třeba věnovat zvýšenou pozornost, aby nedošlo k jejich zneužívání.

Klíčová slova: web 3.0, web, sémantika, sémantický web, mikrodata, mikroformáty, RDF, linked data, XML, Rich Snippets

Summary

The aim of the diploma thesis is to outline the importance of the concept of web 3.0. Define its individual properties with a focus on the most important aspect, the process of semantic interpretation. The work deals with the ways and technologies in which this process is beeing done.

These approaches are then implemented in practice and the results will be to select the optimal tool for the implementation of structured data within a web document.

The work deals with the architecture of the Semantic Web. It analyzes the individual components and the ways in which they are connected into one working model, which allows the subsequent use of advantages of structured data in them.

It is necessary to identify the advantages and threats that the Semantic Web will bring. Find fields, to which is necessary to pay attention to prevent their misuse.

Keywords: web 3.0, web, semantics, semantic web, microdata, microformats, RDF, linked data, XML, Rich Snippets

Obsah

1 Úvod

Web 3.0 je pojem, který se obtížně identifikuje. Nemá vlastní definici a je používán v mnoha oborech. Ačkoliv první zmínky o něm pochází z dalšího nástinu rámce vývoje technologie zvané jako World Wide Web[1], tak dnes se hojně využívá v marketingu jako nástroj pro obchodníky jak zviditelnit své produkty. V roce 2012 byl dokonce vyhlášen na 93. místě v žebříčku nejvyhledávanějších „buzzwords" serveru marketing-jive.com[2], tedy slov vyjadřujících nové, revoluční, módní termíny, které slouží k získání pozornosti či dokonce ohromení druhého. Lze ale říci, že se jedná o výsledek snahy charakterizovat průběh vývoje webu.

Toto rozdělení na pomyslné etapy přišlo společně s novými funkcemi, které web obsahoval. V první polovině první dekády 21. století došlo ke změně konceptu pevných, statických webových stránek na dynamickou platformu určenou ke sdílení a společnou tvorbu obsahu. Když začali být rozdíly mezi funkcemi, které měl web ve svém počátku a funkcemi, kterými v té době začal disponovat, opravdu markantní, tak se začal označovat jako svá druhá generace. Vznikl pojem web 2.0 a zpětně jako retronym se objevilo i pojmenování web 1.0 pro označení první generaci webu jako takového.

Podobně se v poslední době začaly objevovat snahy pojmenovat nový směr vývoje za účelem změnit data na strojově čitelné, strukturované informace. Toto umožní mimo jiné i porozumění významu dat či dotazů se kterými uživatel pracuje. Představme si, že vyhledává komedii, na kterou by šel do kina a poté do čínské restaurace. Což znamená, že bude vyhledávat kina v okolí, poté přejde na jejich stránky, zjistí, kde se přesně nachází, najde si seznam filmů, vyfiltruje si komedie a tak dále. Také ovšem nechce, aby to poté z kina nebylo moc daleko, musí tedy ten samý proces opakovat s restauracemi, porovnává jejich vzdálenosti a celkem tak navštíví tucty stránek, než teprve bude moci vůbec vyjít ze dveří.

Web 3.0 by měl právě takovéto úlohy velice zjednodušit. Prohlížeč by již dokázal sám najít seznam restaurací, kin i jejich filmů a podstatně by tento proces zjednodušil. Toto

[1]World Wide Web, nebo-li zkráceně WWW či web.Aplikace na síti propojených počítačů přes protokol TCP/IP za účelem usnadnění sdílení informací. Jedná se o označení pro celosvětovou síť hypertextových dokumentů, které jsou prezentovány internetovým protokolem http. V práci je pro označení této služby využíván výraz „web".
[2]http://www.marketing-jive.com/2012/01/top-100-marketing-buzzwords-for-2012.html

je možné v případě, že by prohlížeč dokázal porozumět obsahu stránek a dokázal tak identifikovat jednotlivé entity na stránce. Nebude již na obsah stránky pohlížet jako sérii alfanumerických řetězců, tedy stringů, ale na slova, která mají vlastní význam. Prohlížeč pak zanalyzuje vaši otázku a zobrazí vám možné výsledky hledání.

To ovšem nebude vše. Kromě automatického vyhledávače, který zkoumá výskyt hledaného shluku znaků (vyhledávaný výraz) a dle jejich četnosti seřazuje výsledky hledání, by se měl nově web 3.0 začít chovat jako jakýsi osobní asistent. Během doby, kdy uživatel „surfuje“ na internetu, se prohlížeč učí, co jej zajímá, v závislosti dle obsahu stránek a aktivity, které na nich vyvíjí. Čím více internet tak využívá, tím více toho o něm prohlížeč ví a tím méně musí být specifický ve své otázce. Nakonec by mohli jednoduše jen napsat do vyhledávače „Kam mám jít se najíst?“ a prohlížeč už automaticky na základě uživatelovi historie zjistí, které restaurace by byli vhodné. Kdy dal tlačítko „like“ nebo jim přiřadil vysoké honocení. K tomu přidá do výpočtu uživatelovu momentální polohu a je tak schopen rovnou navrhnout seznam restaurací, které by byli pro uživatele vhodné.

Je však třeba analyzovat jakým způsobem je zajištěna funkčnost strojové čitelnosti významu jednotlivých entit uvnitř dokumentu a které technologie se na této činnosti podílejí.

2 Cíl práce a metodika

2.1 Cíl práce

Cílem práce je poskytnout ucelený pohled na vývoj webových prezentací v souvislosti s rozvojem technologií pro značkování sémantického významu objektů uvnitř webových dokumentů. Základem je zmapování vývoje termínu označovaného jako Web 2.0 a analýza složek, které určují jeho další směřování. Výsledkem práce je zhodnocení následného rozvoje do stavu, který lze označit jako Web 3.0.

2.2 Metodika

Prvním krokem je studium odborné literatury, webových prezentací a odborných diskuzí věnujících se problematice Webu 2.0 a jeho dalšího vývoje. Získané informace budou syntetizovány a výstup bude použit jako podklad pro analýzu technologií, které jsou uvažovány jako prostředek pro transformaci do stavu označovaného jako Web 3.0. Jejich analýzou a následnou implementací v praxi bude vyslovena hypotéza o optimálních nástrojích pro jeho implementaci a zajištění funkčnosti. Zároveň budou podrobeny zkoumání přínosy i eventuální hrozby tohoto řešení. Výstupem bude zhodnocení, zda již bude možné v dohledné době používat termín Web 3.0 a jaké bude mít dopady na elektronickou prezentaci dat.

3 Teoretická část

V teoretické části práce je představen vývoj webu ve smyslu zaměření této práce a směr jakým se předpokládá jeho další postup do budoucnosti, zejména po transformace na přiřazení ontologických slovníků jednotlivým elementům a popisu architektury služby, která by se dala nazvat jako sémantický web.

3.1 Vývoj webu

Následující kapitola se zabývá vývojem, kterým služba World Wide Web prošla od svého vzniku. Ten je rozdělen na jednotlivé etapy v závislosti na charakteru směru vývoje webu v daném období. Je to z důvodu názornosti a lepší přehlednosti, nejedná se o oficiální rozdělení.

3.1.1 Vznik webu

Vznik počítačů a jejich rozšíření po celém světě znamenal potřebu decentralizované sítě uzlů schopných vzájemné komunikace. Prvním velkým impulzem byla studená válka (zhruba 1947-1991) mezi USA a dnes již bývalým Sovětským svazem. Obě mocnosti byly vyzbrojeny zbraněmi hromadného ničení umožňujícími masivní ničení infrastruktury a byla tak potřeba eliminovat v rámci sítě možné řídící uzly podobné telefonním ústřednám a zabezpečit tak, aby jednotlivé uzly byli při přerušení navzájem nezávislé. Americká agentura Advanced Research Project Agency (ARPA) pod záštitou Ministerstva obrany USA tak v roce 1969 představila síť, kterou nazvala ARPANET. Ta nejprve sloužila jako armádní síť spojující výzkumná a vojenská centra. Postupem času se ovšem k této síti začaly připojovat i univerzity a s ní i veřejnost. Došlo tak k odklonu armádní části do vlastní větve, pojmenované jako Milnet. Původní síť ovšem zůstala dále udržovaná americkou armádou a různými vládními agentury. Tehdy vzhledem k malému, úzce zaměřenému kruhu uživatelů, nenabízel komerční využití. Zásadní problém spočíval v nekompatibilnosti jednotlivých platforem používaných v různých zařízení připojených k síti. Tento problém byl ovšem vyřešen v roce 1983 představením protokolu TCP/IP,

protokolu který sjednotil způsob, jakým probíhala výměna dat mezi různými uzly. Tato síť je dnes známá jako internet.

Sjednocení komunikačních protokolů umožnilo v roce 1989 vznik nové aplikace pro sdílení dat. Jejím autorem se stal Tim Berners-Lee, který v CERN[3]upracoval jako programátor. Přišel s nápadem univerzální služby pro přenos hypertextu, které jinak byly do té doby nesjednocené a na různých zařízení se různě zobrazovaly. Ačkoliv v té době již existovaly jazyky na jeho formátování, tak nebyly sjednocené a na každou možnost úpravy text byly používány rozdílné způsoby. Toto bylo vyřešeno vytvořením HTTP[4] protokolu a k němu prohlížeče WorldWideWeb na tehdejším operačním systému NextStep. Později ovšem tento prohlížeč přejmenoval na Fortrans, aby nedocházelo k záměně názvu prohlížeč s nově vzniklou službou. Ta vznikla jako reakce na potřebu zjednodušení sdílení dokumentů, které vznikají na různých místech a byly k dispozici již v hypertextové podobě. Tento systém funguje na principu WYSIWYG[5].

Vzhledem k celosvětovému využití této sítě je potřebná jistá dávka normalizace, technické i softwarové specifikace, aby nedocházelo k nekompatibilitě mezi zařízeními. Toto je předkládáno řadou agentur. Vzhledem k zaměření této práce zmíním W3C – World Wide Web Consortium. To se zabývá normami pro jazyky HTML, XHTML, CSS a DOM (objektový model dokumentu). Tyto jazyky tvoří základ pro dnešní webové prezentace.

3.1.2 Web 1.0

Web of document, nebo jak se také označuje první vývojová etapa webu, se dá jedním slovem shrnout jako statická. Webové prezentace se úzce držela myšlenky WYSIWYG, kdy uživateli byla předkládána data bez jakékoliv možnosti jejich personalizace či úpravy v rámci prohlížeče. Jednalo se o statické dokumenty, jež byly uživateli předkládány v již nepozměnitelné podobě. Jediný způsob, který uživateli

[3] CERN (z franc. Conseil Européen pour la recherche nucléaire) - Evropská laboratoř pro výzkum fyziky částic ležící na francouzsko-švýcarské hranici poblíž Ženevy. Byla založena v roce 1954 a jedná se o nejrozsáhlejší výzkumným centrem částicové fyziky na světě. Je vybaven LHC (Large Hadron Collider), největším urychlovačem částic na světe.
[4] HTTP (Hyper Transfer Protocol) – protokol pro výměnu dokumentů v rámci sítě internet
[5] WYSIWYG – What you see is what u get (česky co vidíš, to dostaneš) – způsob editace dokumntů, kdy je zobrazená verze totožná s výslednou, např. u tisku

umožňoval měnit obsah webové stránky v rámci samotného prohlížeče, byly služby jako např. questbook. Ty sloužily jako návštěvní kniha, kam mohl uživatel sdělovat například své komentáře k obsahu stránky.

Celkový objem dat na webu byl v době existence této etapy nesrovnatelně menší než dnes. Připojení k internetu ještě nebylo tak rozšířené mezi obyvatelstvem, a tudíž měl na konci své etapy „pouze“ kolem 45 miliónů uživatelů. Taktéž obsah samotného webu byl podstatně menší, kolem 250 000 webových stránek. To bylo dáno tím, že chyběly kapacity a možnosti tvorby dalšího obsahu novými uživateli byly značně omezené.

Obsah webu je vytvářen tak převážně jeho vlastníkem a komunita má na něj velmi omezený vliv. Samotná interakce mezi těmito 2 skupinami je velmi omezená, probíhá zpravidla formou questbooků a je velice obtížné ji dosáhnout. Její nabídnutí provází velké nároky na zprovoznění. Také vzhledem k faktu, že obsah může upravovat pouze vlastník, byla aktualizace stránky velice náročná. Mezi komunitou a webovým obsahem tedy existuje pouze jednosměrné spojení, kdy informace proudí směrem z webu k zákazníkovi. Vzhledem k tomu není ani možnost personalizace pro tuto dobu aktuální.

Hlavní nevýhodou této etapy tak byl nedostatek interakce mezi uživatelem a samotnou aplikací a obtížné šíření dalších informací.

3.1.3 Web 2.0

O druhé etapě vývoje webu se mluví od roku 2004, i když samotný termín vznikl už o několik let dříve. Dalo by se říct, že pokud o první etapě mluvíme jako o jednosměrné komunikaci mezi webem a uživatelem, zde již dochází k obousměrnému toku informací, kdy sám návštěvník je schopen s obsahem stránky manipulovat a dokonce jej i měnit. Taktéž statická forma informací byla změněna a web se stal dynamickým, sociálním a přizpůsobivým.

Na rozdíl od první etapy se zde již zapojuje celá komunita a tvorba webu se tak stává kolektivní činností, kdy se na tvorby obsahu dané stránky podílí celé komunita. To je dáno rozvojem nástrojů umožňující například přidávání článků, hodnocení, recenzí apod. Tato interakce mezi čtenáři a vlastníkem webu se stala vítaným jevem. Má formu

diskuzí, chatu, propojení se sociálními sítěmi a podobně. Díky těmto vlastnostem se i aktualizace webu stává jednodušší, protože do značné míry to zvládnou i sami návštěvníci, i když samozřejmě záleží na zaměření samotné stránky. Např. novinky.cz, internetový portál zabývající se zpravodajstvím, umožňuje přidávat články nejen redaktory, ale také samotnými uživateli ve formě blogů. Zajímavé je, že se tímto způsobem může sama komunita stát součástí webu a uživatel se tak stane tím, „o kom web píše“.

Důležitým faktorem také bylo, jak již bylo zmíněno v kapitole Úvod, že se webové stránky již přestaly být pouhými izolovanými dokumenty v síti a staly se dynamickými platformami pro externí aplikace koncovému uživateli, pomocí kterých mohl uživatel komunikovat.

3.1.4 Web 3.0

Pojem web 3.0 se začal objevovat kolem roku 2006, kdy začal být zmiňován mnoha odborníky v oboru. Postupem času čím dál více odborníku přicházeli s myšlenkou nové generace webu jako takového. První byl Jeffrey Zeldman[6], který jej již začátkem roku 2006 zmínil ve své práci, kde se snažil kritizovat web 2.0 a jeho stoupající složitost. Tam o něm ještě psal v lehké ironii v nadsázce nad dalšími vývojovými stupni webu.

Někdy se také prvenství připisuje Johnu Markoffovi[7], který web 3.0 zmínil ve svém článku o inteligentním webu v New York Times ke konci roku 2006. Ten byl sice vydán později než Zeldmanova práce, avšak Markoff zde již o něm předkládá první definice a vlastnosti, které by novému vývojovém článku webu měli patřit. Poprvé zde popsal propojení mezi sémantickým webem. Popisuje web 3.0 jako systém, který při odpovídání na otázku „Která restaurace je nejlepší?“ bere v potaz i komentáře na stránkách jednotlivých restaurací v rámci kognitivní dedukce, které bude zkoumat prostřednictvím softwarových agentů, kteří by měli pátrat na internetu a veškeré informace kategorizovat a značkovat.

Pud web 1.0 propojil obsah odkazy a web 2.0 propojil obsah s lidmi, tak web 3.0 začíná propojovat obsah, lidi a aplikace za použití sémantiky a standardů. Chce tak poskytnout platformu pro přiřazení ontologickým slovníkům jednotlivým prvkům.

[6]ZELDMAN, J., *A List Apart : For People Who Make Websites*
[7]MARKOFF, J., *The New York Times*

Prostřednictvím implementace sémantiky jsou tak výsledky vyhledávání přesnější, relevantnější a smysluplnější. Dokáže popsat jednotlivé vlastnosti mezi elementy obsažené uvnitř zkoumaného dokumentu. Takové síti se pak nazývá Sémantický web, Web of Data, Giant Global Graph, nebo také Web 3.0.

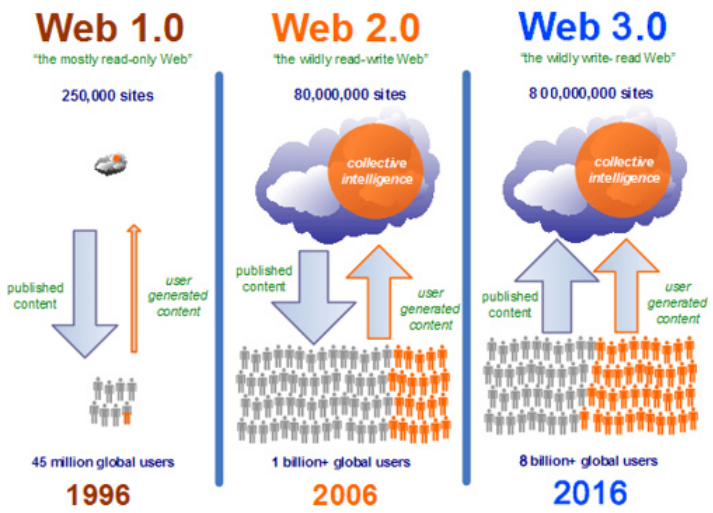
3.1.5 Porovnání

Web v kontextu této práce prošel za dobu své existence mnohými změnami a jeho dnešní podoba podporuje mnohem větší spektrum uživatelových potřeb než na svém počátku. Výsledné srovnání jednotlivých etap vývoje webu z pohledu zaměření práce je vidět v tabulce 1.

	Web 1.0	**Web 2.0**	**Web 3.0**
Obsah	vytvářen převážně vlastníkem a provozovatelem stránky	komunita se aktivně podílí na tvorbě obsahu	uživatel má kontrolu nad způsobem zobrazení a obsahu dat, obsah strojově čitelný
Interakce & Komunikace	obtížná vzhledem k vysokým nárokům na provozovatele stránky, jednosměrná interakce, uživatel je pasivní příjemce dat	propojení se sociálními sítěmi a messengery, obousměrná komunikace, uživatel se může stát tím, "o kom web píše"	propojení více systémů mezi sebou
Sémantika	neexistuje	v jednoduché podobě, využívání indexování a jednoduché databáze slovních spojení	výrazy obohaceny o význam, data propojena díky vzájemným definovaným vztahům
Personalizace	neexistence implicitní možnosti personalizace	možnost vytváření a práce s profily uživatelů	data uložena ve strojové podobě a následně prezentována na základě uživatelových požadavků

Tabulka 1: Porovnání etap vývoje webu
(Zdroj: vlastní práce autora)

Společně se vzrůstající dostupností počítačů a softwarových technologií a s nimi spojených výhod či uživatelského komfortu, se web stává čím dál přístupnější pro čím dál více uživatelů. Tím dochází k růstu objemu dat na síti, viz. obrázek 1.



Obrázek 1: Růst objemu dat v rámci etap vývoje webu
(Zdroj: http://rattanasak.jigsawoffice.com/upload/editor-pic/images/web3_3.jpg)

Velký objem dat na síti znamená zároveň i větší nepřehlednost při vyhledávání relevantních informací. Tím dochází ke vzniku potřeby nových technologií umožňujících dosažení lepších výsledků při vyhledávání a práce s daty. Tento směr se nazývá sémantický web.

3.2 Sémantický web

„Zdá se mi o webu, v němž se počítače stanou schopné analýzy všech dat na webu – obsah, odkazy a transakce mezi lidmi a počítači. „Sémantický web“, který toto umožní, se musí ještě objevit, ale až se tak stane, budou ze dne na den mechanismy obchodu, byrokracie a naše každodenní životy zajišťovat stroje hovořící se stroji. „Inteligentní agent“, po kterých lidé věky volali, se konečně zhmotní.“[8]

- *Tim Berners-Lee, 1999*

Využitím sémantického webu by umožnilo změnit koncept celého webu, tak jak ho známe. Ten byl vytvořen pro člověka na již zmíněném principu WYSIWYG. Jeho přeměna a obohacení o metadata, která by umožnila stroji rozumět významu dat, se kterými pracuje, by znamenala příhod úplně nového zážitku z používání této technologie.

Sémantika z hlediska zaměření této práce představuje více vzájemně propojených oborů. Jedná se o způsob jakým je nastřen nový směr vývoje webu, který se postupně čím dál více objevuje v souvislosti s růstem datového objemu na internetu. Společně se vzrůstajícím množstvím zdrojových dat se zvyšují nároky na přesnější identifikaci a relevantnost vyhledávaných dokumentů. Uživatele hledající jednu informaci na internetu si lze tak představit jako člověka hledající jehlu v kupce sena a ta kupka se neustále zvětšuje. Je tedy třeba zdokonalit vyhledávací metody.

Účelem sémantiky je tak na základě definovaných pravidel určit význam obsahu dokumentů na internetu. Představme si případ, kdy uživatel vyhledává na webovém vyhledávači slovní spojení „jablko a hruška“. Vyhledávač bude projíždět jednotlivé dokumenty na webu a dle úrovně výskytu vyhledávaného výrazu s jejich obsahem bude seřazovat výsledky. Tento hledaný výraz je ovšem pro něj stále pouze série znaků, jejichž shodu porovnává. Účelem tak je přidat jednotlivým výrazům jejich význam. Pokud vyhledávač rozumí, co daná slova z našeho příkladu znamenají tak Jablko a hruška již tak

[8] *„I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web - the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize.“* – zdroj: http://readwrite.com/2009/03/12/semantic-web-treasure-trove-for-marketers

přestanou být jen série několika znaků (string), ale získají i tak vlastnosti, například že jsou jedlé, ovoce apod. může upřednostňovat výsledky, které jsou spojeny s ovocem. Toto lze využít v mnoha oborech. Kromě relevantnějších výsledků při vyhledávání to také může znamenat možnost komunikace více aplikací na základě obsahu.

Vyhledávání na sémantickém webu se tak následně děje na základě 3 kroků:

- Automatické přiřazování stránek do dynamických kategorií. Zjištění které téma a zaměření má která stránka, kde se výsledek shodoval. Kam vedou odkazy na dané slovo je první krok ve fázi zjištění jeho významu.
- Vrácení příbuzných výsledků zpátky do vyhledávání s nižší váhou než výraz, který zadal uživatel. Stále přitom probíhá kontrola, zda jsou přijímány lepší výsledky na základě jejich vzájemné podobnosti.
- Využití statistické analýzy při zjištění míry spojení hledaného slova s výsledky. Zde se tak přiřazuje priorita v závislosti na aktuálnosti témata, tj. podle počtu výskytů v závislosti na době. Například při vyhledávání slova „volby“ se momentálně v České Republice objevují témata spojená s nedávnou volbou prezidenta.

V případě webových dokumentů může být sémantika doplněna i přímo uživateli, který pomocí tagů označkuje jednotlivé entity v závislosti na jejich vlastnostech a tím je definuje. Informace jsou tak doplněny o ontologii, tedy definici pojmů a vztahů mezi nimi. Současné weby jsou jen změtí dokumentů, kde je stále složitější nalézt relevantní informaci. Je třeba tak nalézt způsob přiřazení významu jednotlivým částem obsahu, který bude lépe pochopitelný pro počítač i pro člověka.

3.2.1 Postup sémantizace

Krátce poté, co byla myšlenka sémantizace poprvé přednesena na konferenci v Brisbane z úst Tima Berners-Leee, tak s jeho pomocí byl v roce 2000 spuštěn web http://semanticweb.org, který funguje jako platforma pro komunitu kolem sémantického webu. O rok později již sémantickým webem začala oficiálně zajímat i W3C, která spustila

vlastní web na toto téma[9] a ustanovila skupinu pro vývoj jazyka OWL(Web Ontology Language)[10]. Ten byl zveřejněn společně s RDF a RDFS v roce 2004 jako základní standardy pro práci s ontologiemi na webu. Když pak v roce 2008 přibyl i standard SPARQL pro dotazování a RDFa pro ukládání metadat v rámci atributů XHTML znamenalo to etablování základních nástrojů pro přistupování k ontologiím v rámci webového prohlížeče.

Sémantický web se jako pojem u nás začal objevovat počátkem roku 2007. Na své přednášce Web 2.0 vs. Sémantický web jej na konferenci Inforum zmínil Vilém Sklenák[11], vedoucí katedry informačního a znalostního inženýrství na VŠE. Tento pojem deklaroval jako transformaci stávajícího webu do databázové struktury za využití Sémantického webu a 3D navigace. Jako jeho definici použil vzorec:

Web 3.0 = Web 2.0 + Sémantický web

Sklenák tak považuje Web 3.0 za Web 2.0 doplněný o sémantizaci za účelem integrace dat. Podobným způsobem uvažoval Nova Spivack[12], který pro vyjádření Webu 3.0 použil podobný vzorec:

Web 3.0 = Web 2.0 + 1

Přičemž kde „1“ chápe jako software, kterým lidem pomáhá organizovat a vytvářet smysl webového obsahu. Upozorňuje, že důležitost samotných lidí, kteří pomáhají sami organizovat obsah webu pomocí folksonomie[13] a sociálních sítí.

Co se týče dalšího vývoje Sémantického webu, tak Jiří Procházka[14] tvrdí, že má 3 hlavní fáze. Vývoj technologií, publikování informací pomocí těchto technologií a poté vývoj aplikací a automatických agentů pracující s těmito informacemi. První fáze, kdy se vývoj soustředil především na první odvětví, pominula, teď je třeba rozvíjet druhé odvětví, protože bez něj je nemyslitelný rozkvět odvětví třetího, a přitom právě ono je naším cílem, protože přináší zjednodušení našeho každodenního života.

[9]http://www.w3.org/2001/sw/
[10] Viz. kapitola 3.3.5 OWL
[11]SKLENÁK, Vilém. *Inforum*
[12]SPIVACK, Nova. *Minding the Planet*
[13] Pomocný prostředek pro kategorizaci obsahu internetu. Uživatelé sami popisují obsah webu pomocí tagů, které slouží pro jejich utřídění a pozdější vyhledávání.
[14]http://www.zdrojak.cz/clanky/uvod-do-semantickeho-webu/

Dnes se hodně zmiňuje možnost tvorby databází s definicí jednotlivých pojmů, které by fungovali jako slovníky a umožňovali tak externě přidávat vlastnosti dalším objektům. Veliké naděje se také vkládají do jazyka XML, který se používá pro strukturování dat z věcného pohledu a jeho součástí již tak je i soubor pro definici dat.

Zjednodušeně řečeno, schopnost přiřadit jednotlivým elementům jejich význam by umožnilo pracovat s webovým obsahem jako s relační databází. Uživatel by získal veškerý komfort, který práce s relacemi nabízí a také by se zpřesnily výsledky vyhledávání.

Sémantizaci webu však brání několik faktorů, které pocházejí už z původního způsobu, jakým byl web, tak jak ho známe, navržen. Jedná se zejména o to, že byl tvořen jako databáze hypertextových dokumentů, stránek, nikoliv však jako databáze strukturovaných dat. V dnešní době sice velké portály jako wikipedie a podobně vytváří dynamické dokumenty na základě uživatelova požadavku a datové databáze, jedná se ovšem spíše o výjimky. Většina těchto stránek je tak stavěna na architektuře orientované pro snadné prohlížení a prohledávání člověkem, nikoliv však strojem. Značky se spíš orientují na způsob, jakým různé elementy zobrazit, ale už ne definici jejich významu či dokonce jejich katalogizování. Jejich implementace tak znamená více práce bez viditelného přínosu pro koncového uživatele.

3.2.1.1 RSS

Obrázek 2: Logo RSS (Zdroj: http://www2.macleans.ca/wp-content/uploads/2009/03/rss-icon.jpg)

Jedna z prvních služeb fungujících na principu přidávání metadat kvůli označení jednotlivých částí obsahu. Má podobné vlastnosti jako sémantický web a přesto o něm v souvislosti s ním málokdy uslyšíme. Jedná se o RSS. Jedná se o rodinu XML formátů určený pro čtení novinek na webových stránkách. Tato technologie pochází z roku 1999 a byla vyvinuta společností Netscape. Byl založen na RDF a její zkratka je často interpretována různě (RDF Site Summary nebo také Rich Site Summary). Používá se pro předávání novinek mezi servery. Pokud tak kupříkladu navštěvujeme více zpravodajských serverů a chceme zjistit, kde se objevily jaké články, tak již nemusíme postupně projít

každý jediný zpravodajský portál a vlastnoručně je zkontrolovat. Díky RSS kanálům jsou k dispozici seznamy článků s jejich popisem.

RSS umožňuje přiřazovat článkům v rámci webového dokumentu podobné sémantické vlastnosti, díky nimž poté můžeme s obsahem pracovat samostatně a odfiltrovat ostatní prvky, které se na stránce objevují. Jako příklad uvádím příklad kódování v dnes zřejmě nejpoužívanější verzi RSS 2.0.

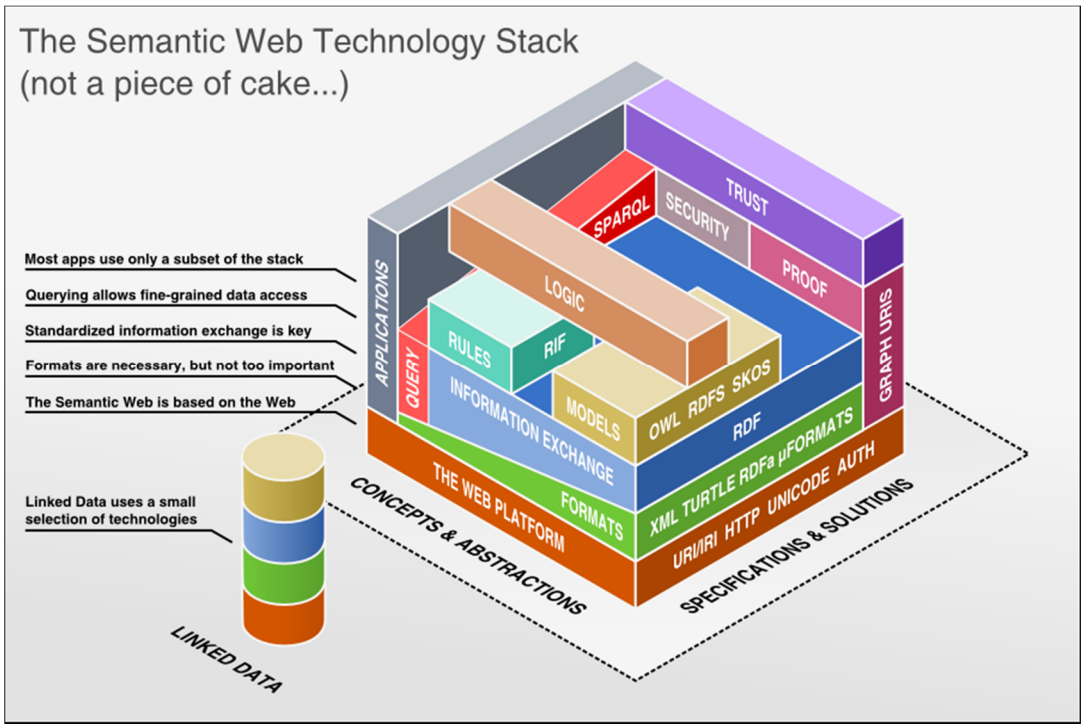
```
<rss version="2.0">
<channel>
<image>
<link>/</link>
<title>RSS</title>
<url>http://kyokushinkai-uray.ru/wp-content/themes/primary/images/rss-
ico.gif</url>
</image>

<title>Diplomová práce </title>
<link>www.example.com</link>
<description>Sémantický web</description>
<language>cs</language>
<pubDate>Ne, 10.3.2013 22:00:05 GMT</pubDate>

<item>
<title>RSS</title>
<link>www.example.com/rss</link>
<description>Tato část je věnována popisu služby RSS.</description>
<author>David Macháček</author>
</item>
  </channel>
</rss>
```

Jak vidíme, tak obsah stránky je rozdělen do bloků pomocí značek jako „description“ či „title“, které již ale neslouží k formátování, nýbrž k definování obsahu textu.

Tento způsob zápisu RSS je velice podobný dnešním technologiím, které se sémantizací zabývají. Ačkoliv ovšem používají podobný způsob, tak se liší zejména v učení dané technologie. Nelze tak říct, že RSS je technologie určená pro přeměnu klasického webu na sémantický. Ovšem vzhledem k velikým podobnostem považuji za důležité ji v této práci zmínit.

3.2.2 Vrstvy sémantického webu

Sémantický web je tak způsob jakým se pracuje s ontologiemi v rámci webu, tento směr je realizován pomocí vzájemně propojených technologií. Ty společně vytváří fungující platformu, kterou bychom nazvali sémantický web. Jejich rozložené je vidět na obrázku níže.



Obrázek 3: Vrstvy sémantického webu
(Zdroj: http://bnode.org/media/2009/07/08/semantic_web_technology_stack.png)

Obrázek ilustruje strukturu hierarchie sémantického webu, kde každá vrstva maximálně využívá schopnosti vrstvy pod sebou. Toto schéma je známé jako The Semantic Web Stack, Semantic Web Cake, či někdy též Semantic Web Layer Cake. Zobrazuje způsob, jakým pomocí zmíněných technologií dosahuje myšlenky sémantického webu. Také jej předurčuje nikoliv jako konkurenta nebo následovníka současného webu, ale jako jeho rozšíření.

Jednotlivé vrstvy se postupně snaží vyřešit a standardizovat konsorcium W3C. Technologie, které schéma zobrazuje od spodní vrstvy až po OWL jsou v současné době již schválené a standardizované pro využití pro tvorbu sémantického webu. Avšak není

ještě jasné, jaké vrstvy přijdou na vrch schémata, jelikož je potřeba implementace všech vrstev pro dosažení sémantického webu.

Tyto vrstvy lze také rozdělit do 3 hlavních skupin:

Technologie Hypertextového Webu

- URI/IRI (Unique/Internationalized Resource Identifier)
- Unicode
- XML + jmenné prostory + schéma

Technologie Sémantického Webu

- RDF (Resource Description Famework)
- RDF Schema
- OWL SPARQL
- RIF

Neidentifikované Technologie Sémantického Webu

- Šifrování
- Uživatelské prostředí
- Důvěra

3.3 Technologie sémantického webu

Následující kapitola se zabývá po detailnější stránce jednotlivými technologiemi, které dohromady tvoří architekturu sémantického webu.

3.3.1 URI/IRI a Unicode

Jedná se o základní kameny každé datové struktury, tedy způsob její adresace, který umožňuje jednoznačnou identifikaci elementů v rámci datové struktury a také určení znakové řeči schopné interpretovat požadovaný jazyk.

URI/IRI je anglická zkratka pro Unique Resource Identifier, resp. Internationalized Resource Identifier. Poskytuje unikátní identifikaci zdrojových dat, to je zejména potřeba kvůli vrchním vrstvám, kde unikátní identifikace je důležitá pro manipulaci s daty. URI využívá pouze znaky ASCII, kdežto IRI využívá jakoukoli znakovou sadu Unicode. Používání znakové sady Unicode a identifikátorů v podobě IRI je dnes již běžnou ustálenou praxí, kterou programátoři akceptují a obvykle dodržují.

Unicode je tabulka znaků všech existující abeced. Obsahuje více než 110 000 znaků a slouží pro práci s textem ve více jazycích. Sémantický web by měl být přístupný všem, bez ohledu na jazyk a písmo, které používají, proto se musí využívat celou sadu znaků Unicode.

Společně s http tak zajišťují připojení na webovou platformu.

3.3.2 XML

(X)HTML bylo vyvinuto k zobrazení dat, naproti tomu, XML slouží k jejich uchovávání a transportu. Jedná se o značkovací jazyk, který byl vyvinut a je spravován konsorciem W3C. Vychází ze SGML jako jeho jednodušší forma.

Jazyk je určen především pro výměnu dat mezi aplikacemi a pro publikování dokumentů. Popisuje strukturu dat z věcného hlediska a již se nezabývá vzhledem.

Pokud si dnes uživatelé posílají mezi sebou data, je potřeba je zasílat ve tvaru, formátu, který bude pro obě strany snadno čitelný. Formáty jako .doc, .xls apod. vyžadují speciální software, který dále většinou vyžaduje jistý operační systém či platformu, který je potřeba pro přečtení těchto formátů. Nelze však předpokládat pro globální sdílení dat, že každý uživatel bude mít přístup ke stejnému softwaru. Vznikla tak potřeba vytvořit jednoduchý otevřený formát, který nebude vázán na žádnou platformu či operační systém. Tím by měl být právě XML. Jeho specifikace je díky vedení W3C zdarma přístupné všem. Každý tak může do svých aplikací lehce implementovat podporu XML. Jedná se o velikou výhodu vůči ostatním např. různým firemním formátům, k nimž není k dispozici povětšinou žádná dokumentace a jedná se o složité, často binární, formáty.

Další velkou výhodou XML je kromě snadné implementace a nezávislosti na platformě už samotný fakt, že byl od počátku vyvíjen jako univerzální jazyk pro uchovávání dat. Vzhledem k tomu, že s ním již bylo od počátku počítáno jako s globální technologií, není úzce spjat s anglickým jazykem, jako většina jazyků pro práci s daty (různé příkazy mají ekvivalent v angličtině, např. for). Jako znaková sada se tak implicitně používá Unicode[15], což umožňuje vytvářet dokumenty v mnoha jazycích. Znamená to také, že XML neobsahuje předdefinované tagy. Je třeba si tak definovat své vlastní. Můžeme tak tvořit jakékoliv značky. Značky budou následně interpretovány dle souboru DTD (Document Type Definition).

3.3.2.1 DTD

Jedná se o doprovodný jazyk, který slouží jako definice jednotlivých tagů, které si uživatel při tvorbě kódu nadefinuje. Struktura DTD funguje jako databáze jednotlivých značek, elementů i atributů společně s jejich popisem. Díky tomu mohou být používány uvnitř XML jako klasické značky, tagy.

DTD vzniklo společně s XML.

[15] Unicode je tabulka znaků všech existujících abeced, která v současnosti obsahuje více než 110 000 znaků. Unicode umožňuje pracovat se znaky všech písem i různými jinými symboly stejným způsobem, takže mohou být využívány současně.

3.3.2.2 XML Schéma – XSD (XML Schema Definition)

Velká výhoda XML spočívá v možnosti používání vlastní sady značek pro přidání sémantiky do dokumentů XML. Tyto značkovací jazyky je ale třeba definovat a z toho důvodu XML standard přímo obsahuje DTD, kde jsou jejich popisy obsaženy. Tento jazyk pro popis schématu dokumentu pochází ještě z jazyka SGML[16]. To znamená, že je podporován velkým množstvím aplikací. Oproti XSD má ovšem 3 základní nevýhody:

- Neexistence podpory jmenných prostorů – umožňuje implementace více sad značek do dokumentu XML, např. vektorový obrázek SVG či rovnici MathML[17]
- Nemožnost určit datový typ – SGML a později tedy i XML bylo vyvinuto pro práci s textovými dokumenty jako knihy, technická dokumentace či webové stránky. Dnes se však XML masově využívá i pro výměnu strukturovaných dat mezi informačními systémy, kde je vhodné definovat obor hodnot pro jednotlivé typy dat, např. faktury. Je tedy potřeba definovat, kromě struktury elementu, také jeho obsah
- Syntaxe DTD se jinde než v XML nepoužívá, ostatní jazyky pro popis schématu dokumentu, které přišly po DTD, používají pro zápis schématu přímo XSD. Zápis je sice složitější, ale umožňuje následně stejný komfort práce jako s XML dokumentem, např. pomocí XSLT[18] je možné generovat dokumentaci či použití CASE[19] nástrojů pro vytvoření datového modelu

Předchozí 3 důvody zapříčinily již v době vzniku standardu XML snahu vyvinout řešení, které by překonávalo daná omezení. Významnou roli při jejich vývoji sehrál zejména jazyk XML-Data, který byl jako návrh zaslán v roce 1998 do W3C. Šlo o jazyk, který umožňoval definici tříd a přitom nebyl striktně vázaný pouze na XML. Posléze došlo k jeho vývoji do zkrácené formy pro jednodušší implementaci jménem XDR (XML Data

[16] SGML (Standard Generalized Markup Language – univerzální značkovací metajazyk, který umožňuje definovat značkovací jakzyky jako své podmnožiny
[17] Viz. Kapitola XML Jmenné prostory
[18] XSLT (eXtensible Stylesheet Language Transformation) slouží k převodům dat ve formátu XML do libovolného dalšího formátu, nejčastěji HTML
[19] CASE (Computer-Aided Software Engeneering) – využití softwaru pro vývoj a tvorbu nových aplikací

Reduced), který se postupně stal novým standardem, až byl v roce 2001 oficiálně doporučen W3C jako XML Schéma.

3.3.2.3 Implementace

XML Schema a DTD jsou tak nejčastějšími standardy, které popisují schéma XML dokumentu. V praxi rozdíl mezi těmito službami vypadá následovně:

XML

```
<?xml version="1.0" encoding="utf-8"?>
<zamestnanci>
<zamestnanec id="101">
<jmeno>Jan</jmeno>
<prijmeni>Novák</prijmeni>
<email>jan@novak.cz</email>
<email>jan.novak@firma.cz</email>
<plat>25000</plat>
<narozen>1965-12-24</narozen>
</zamestnanec>
<zamestnanec id="102">
<jmeno>Petra</jmeno>
<prijmeni>Procházková</prijmeni>
<email>prochazkovap@firma.cz</email>
<plat>27500</plat>
<narozen>1974-13-21</narozen>
</zamestnanec>
</zamestnanci>
```

Definice elementů použitých v XML dokumentu tak lze zapsat dle uvážení dvěma variantami - DTD a XSD:

DTD

```
<!ELEMENT zamestnanci (zamestnanec+)>
<!ELEMENT zamestnanec (jmeno, prijmeni, email+,
                plat?, narozen)>
<!ELEMENT jmeno      (#PCDATA)>
<!ELEMENT prijmeni   (#PCDATA)>
<!ELEMENT email      (#PCDATA)>
<!ELEMENT plat       (#PCDATA)>
<!ELEMENT narozen    (#PCDATA)>
```

```
<!ATTLIST zamestnanec
        id     CDATA #REQUIRED>
```

XLM Schema (XSD)

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
<xs:element name="zamestnanci">
<xs:complexType>
<xs:sequence>
<xs:element name="zamestnanec"
                maxOccurs="unbounded">
<xs:complexType>
<xs:sequence>
<xs:element name="jmeno" type="xs:string"/>
<xs:element name="prijmeni" type="xs:string"/>
<xs:element name="email" type="xs:string"
                    maxOccurs="unbounded"/>
<xs:element name="plat" type="xs:decimal"
                    minOccurs="0"/>
<xs:element name="narozen" type="xs:date"/>
</xs:sequence>
<xs:attribute name="id" type="xs:int"
                    use="required"/>
</xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema> Schema
```

3.3.2.4 XML jmenné prostory

XML dokument v sobě může obsahovat více elementů a atributů, které jsou využívány různými softwarovými moduly, které ovšem potřebují umět rozpoznat, které objekty v rámci dokumenty jsou určeny jim ke zpracování. K zajištění této modularity je potřeba najít způsob, jakým označkovat jednotlivé elementy pomocí vlastní URI. Tato adresa je popsána v tzv. rozšířeném názvu, který se skládá jak ze jména daného jmenného prostoru popsaného pomocí URI a jeho lokálního názvu pro snazší identifikaci.

Využitím jmenných prostorů je umožněno použití více identicky pojmenovaných objektů na jedné stránce. Odstraňuje tak konflikty, které vznikají při snaze rozpoznat je od sebe.

3.3.3 RDF

Poté, co byla známa potřeba obohatit data o metadata, tudíž přiřadit jim význam, vznikaly nové systémy, které se o to snažily. Organizace W3C proto definovala způsoby pro zachycení sémantické informace objektů a vztahy mezi nimi pomocí standardu RDF (Resource Data Framework). Jedná se o obecný rámec pro popis, výměnu a znovupoužití metadat.

RDF datový model je postaven na stejném principu jako modelování v E-R diagramu či class diagramu. Funguje na principu tvorby výroků ve formě Objekt – Předmět – Vlastnost (objekt - subjekt - predikát). RDF terminologie pro tento princip tvorby výroků zavádí termín trojice neboli triples. Objekt určuje, o jaký zdroj jde, vlastnost zase charakter vztahu a předmět ukazuje na objekt či hodnotu vlastnosti. Například výrok „Obloha má modrou barvu.“ je v RDF tvořena jako trojic řetězce: objekt je „obloha“, predikát „má barvu“ a subjekt je „modrá“. Toto umožňuje i zpětný postup, kdy na základě těchto znalostí lze i zpětně odvodit, že modrá je obloha. Je třeba ještě dodat, že objektem může být vedle řetězce znaků také vnější zdroj. Jedná se o objekt, kterému je přiřazen jednoznačný identifikátor URI.

Obrázek 4: Struktura RDF grafu
(Zdroj: http://www.ikaros.cz/images/201110/motejlkova1.png)

Tímto způsobem se popisují vlastnosti subjektu, jeho hierarchii apod. Význam slov je získáván z ontologií, tedy RDF slovníků, jejichž struktura je popsána pomocí RDF schéma. Tyto slovníky jsou decentralizované a mohou se vzájemně kombinovat.

Implementací trojic lze tedy definovat význam jednotlivých elementů. Je to tedy jedna z cest, jak dosáhnout sémantického webu. Standard tak umožňuje slučování nesourodých dat do strukturované databáze objektů s různými vlastnostmi. Data jsou tak čitelná na různých zařízení při zachování jejich významu.

3.3.3.1 Serializace

Způsob zapisování dat přes RDF je realizováno 2 hlavními způsoby. První z nich využívá pro zápis formát XML a je nazýván tak RDF/XML. Ačkoliv je tento způsob snadno strojově čitelný, je obtížný na syntaxi a logickou strukturu. Z toho důvodu byl vyvinut další způsob, který již XML nevyužívá a byl vytvořen jako pro člověka lépe čitelná verze RDF/XML jménem Notation 3 (N3). Jedná se o kompaktnější a lépe pro člověka čitelnou verzi RDF/XML.

Z hlediska zaměření této práce je důležitá jeho podmnožina, zvaná Turtle, která slouží k implementaci RDF trojic. N-Triples je dalším krokem směrem k simplifikaci způsobu syntaxe a jejího dalšího zkrácení a zjednodušení. Jeden řádek N-triples vyjadřuje jednu trojici, a ačkoliv je tento zápis delší než zápis v Turtle, je podstatně uživatelsky přívětivější při tvorbě většího množství řádků.

V následujícím příkladu je popsán způsob vytvoření RDF trojicí, které vyjadřují název stránky v RDF/XML, Turtle a N-Triple.

RDF/XML

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:dc="http://purl.org/dc/elements/1.1/">
<rdf:Description rdf:about="http://www.xmacd108.php5.cz/">
<dc:title>Web 3.0</dc:title>
</rdf:Description>
</rdf:RDF>
```

Turtle

```
@prefix dc: <http://purl.org/dc/elements/1.1/>.
<http://www.xmacd108.php5.cz/>
  dc:title "Web 3.0";
```

N-Triple

<http://www.xmacd108.php5.cz/><http://purl.org/dc/terms/title>„Web3.0“.

3.3.3.2 RDF Schema

RDF Schema (RDFS) je první jednoduchý sémantický jazyk využívající koncepci RDF. Vznikl v roce 1999 pod záštitou konsorcia W3C. Umožňuje specifikovat vlastnosti jednotlivých objektů a jejich hodnoty. Naplňuje tak požadavky pro vkládání ontologie do obsahu webových stránek vytvořením slovníku zapsaného ve standardu RDF.

Ačkoliv neobsahuje datové typy, je možné pro popsání vztahů použít následující zápisy.

- rdfs:Class – třída (subClass podtřída)
- rdf:Property – vlastnost (subProperty podtřída)
- rdfs:range – definice oboru hodnot
- rdfs:domain – omezení domény, definičního oboru

3.3.4 SPARQL

SPARQL je dotazovací jazyk pro databáze, který je schopen manipulovat a získávat data uložená ve formátu RDF. Spadá pod skupinu RDF Data Access Working Group (DAWG) patřící do konsorcia W3C. Jedná se o jednu z klíčových technologií sémantického webu. 15. ledna 2008 byla tato technologie přijata jako oficiální doporučení konsorcia W3C ve verzi SPARQL 1.0.

Jedná se o podobný jazyk jako SQL, jelikož dokáže pracovat s objekty, resp. instancemi třídy, které mají různé vlastnosti. Tyto vlastnosti jsou definovány trojicemi[20].

SPARQL zná 4 druhy dotazů:

- **Select** – na získání hodnot ve formátu tabulky
- **Construct** – tvorba dalších prvků ve formátu RDF
- **Ask** – dotaz ve formátu true/false
- **Describe** – výstup ve formátu graf

Příklad na získání odpovědi na otázku „Vyjmenuj hlavní města zemí v Africe“:

```
SELECT ?capital ?country
WHERE {
  ?x abc:cityname ?capital ;
     abc:isCapitalOf ?y .
  ?y abc:countryname ?country ;
     abc:isInContinent abc:Africa .
}
```

Pro doplnění je třeba zmínit ještě další alternativy k SPARQL, jedná se například o MQL, rdfDB, RDQL či SeRQL.

3.3.5 OWL

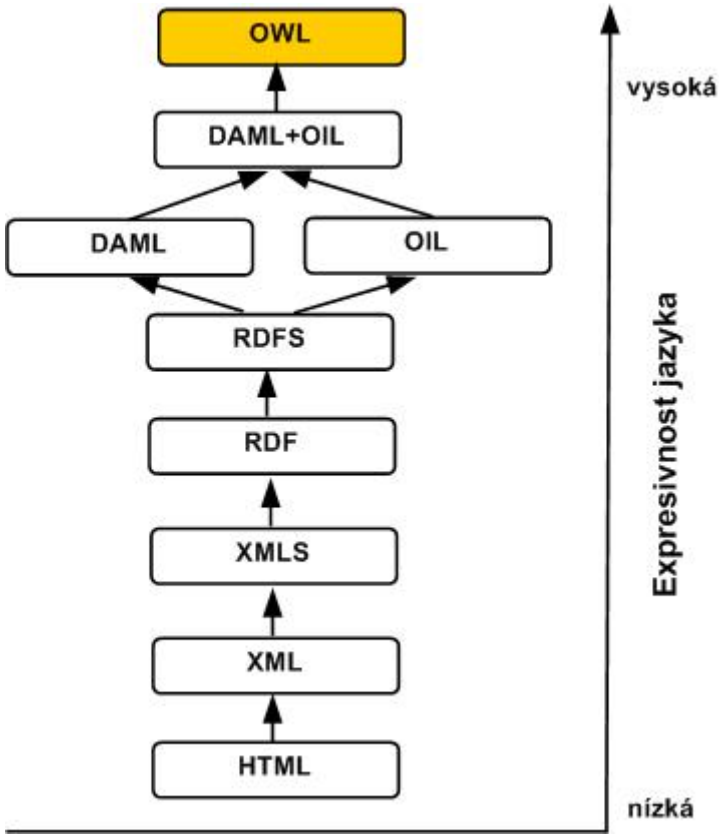
Jedná se o značkovací jazyk vyvinutý organizací W3C (World Wide Web Consorcium) pro tvorbu ontologií právě v prostředí sémantického webu. Jazyk OWL se vyvinul z kombinace jazyků DAML a OIL. Původním záměrem bylo dosažení větší míry expresivity jazyka a tím i bohatší sémantiky, při zachování kompatibility s jazyky XML a RDF. Zároveň měl umožňovat i odvozování nových skutečností z již existujících. Výsledkem byl standard vydaný jako oficiální doporučení W3C jménem OWL pro tvorbu nových ontologií 10. 3. 2004.

Většina dokumentů na webu je stále ve formátu HTML. HTML se spíše zaměřuje na způsob zobrazení webové stránky pro člověka než na strojovou čitelnost. XML jazyk

[20] Viz. kapitola 3.3.3 RDF

se snaží tyto nedostatky HTML vyřešit pomocí tagů a dávat tak vzniknut XHTML. Tyto tagy teprve přiřadí dané informaci význam, o kterém se na dané stránce hovoří. Ovšem i XML má své nedostatky, díky nimž nedokáže detailně vystihnout vztahy mezi objekty a lépe je tak popsat. Právě jazyk DAML dokázal odvozovat výroky a stal se rozšířením verzí XML a RDF. Z tvrzení: „Trojúhelník je geometrický útvar“ dokázal odvodit, že například každý trojúhelníkový útvar je zároveň i geometrickým tvarem.

Naopak OIL (Ontology Interference Language) je ontologickou infrastrukturou pro sémantický web založenou na deskriptivní logice, s níž právě OWL pracuje. Po stránce syntaxe čerpá z XML a XMLS a je rozšířením k RDF a RDFS.



Obrázek 5: Porovnání expresivity jednotlivých jazyků
(Zdroj: http://lide.uhk.cz/fim/ucitel/fshusam2/lekarnicky/zt1/zt1_obrazky/kap04/owlhistorie.jpg)

OWL jazyk se distribuuje ve 3 verzích, OWL-Lite, OWL-DL a OWL-Full:

3.3.5.1 OWL-Lite

Syntakticky nejjednodušší verze OWL. OWL je používána v případě potřeby méně strukturované ontologie s jednoduchým omezením. Je vhodná také v případě, kdy je potřeba již existující databáze XML a RDF rozšířit o OWL reprezentaci. Jedná se tak o podmnožinu jazyka OWL-DL. Na jazyk byla aplikována další omezení, která snižují jeho vyjadřovací schopnost. Jazyk je pak také jednodušší a efektivnější na zpracovávání.

3.3.5.2 OWL-DL

OWL-DL je kompromisem mezi výpočetní výkonností a vyjadřovací silou. Je složitější než OWL-Lite a je založena na deskriptivní logice (odtud právě DL). S touto verzí je možné realizovat odvozování a určit tak ontologickou hierarchii, případně zkontrolovat konzistentnost nebo tvorbu komplexnějších popisů a definicí tříd. Obsahuje všechny konstrukce OWL-Full, ale jejich použití je omezené. Není možné např. definovat stejně pojmenované různé druhy dat (např. používat stejné jméno pro atribut a třídu).

OWL-DL společně s OWL-Full rozšiřují expresivnost jazyka o nové výrazy jako hasValue (odkazuje na hodnotu jiného atributu), oneOf (výčet objektů s určenou vlastností) a podobně.

3.3.5.3 OWL-Full

Obsahuje všechny konstrukce a elementy jazyka. Nemá žádné omezení, ale odvozování s ní je složité vzhledem k omezení současných DL odvozovačů (klasifikátorů). Umožňuje libovolně kombinovat s výrazy RDF a RDFS. Umožňuje také měnit význam výrazů v OWL a RDF aplikováním výrazů navzájem. Výhodou této varianty je zpětná sémantická i syntaktická kompatibilita s RDF. Každý RDF dokument je tak také dokumentem OWL, stejně i tak závěr vyvozený na základě jazyka RDF je tak i závěrem v jazyce OWL. Z toho plyne, že OWL-Full Na druhou stran ovšem vysoká složitost jazyka vede k nemožnosti úplné výpočetní podpory pro odvozování a vysoké složitosti zpracování jazyka.

Každý z těchto jazyků je tak pouze rozšířením svého předchůdce, lze tak potvrdit pravdivost následujících výroku (ale již nikoliv zpětně). Každá platná OWL Lite ontologie je platná OWL-DL ontologie, stejně tak každá platná ontologie OWL-DL bude platná i v OWL Full ontologii. To samé lze říct i o výsledcích. Naopak je důležité vědět, že každý OWL (Lite, DL, Full) je i RDF dokumentem, ale naopak to platí pouze pro OWL Full. Vývojáři ontologií by tak sami měli zvážit, který jazyk je pro jejich účely nejvhodnější.

Pro doplnění o další podobné ontologické jazyky je třeba zmínit ještě například Cyc, Ontolingua, OKBC, XOL, SHOE či Ontobroker.

3.3.5.3.1 SHOE

SHOE neboli Simple HTML Ontology Extension byl první jazyk, který sloužil jako rozšíření (X)HTML o sémantické prvky pomocí zpřístupnění elementů jako třída, podtřída či vlastnické vztahy. Jeho největší nevýhodou byla vazba objektu na konkrétní (X)HTML stránku pomocí URL. Byl vyvinut kolem roku 1996 na Marylandské univerzitě.

3.3.5.3.2 Ontobroker

Vznikl v podobné době jako SHOE na univerzitě v Karlsruhe. Má podobnou stavbu, ale důslednou centralizace předpokládající existenci centrálního serveru.

3.3.5.3.3 DAML+OIL

DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer) byl vyvíjen od roku 2000 jako projekt DAML americkou agenturou DARPA, postupně ovšem v zájmu lepších výsledků zkombinován s OIL.. Jeho cílem je předložit sémantický jazyk s větší vyjadřovací silou než má RDF Schema. Postupně však byl překonán OWL.

3.3.6 RFI

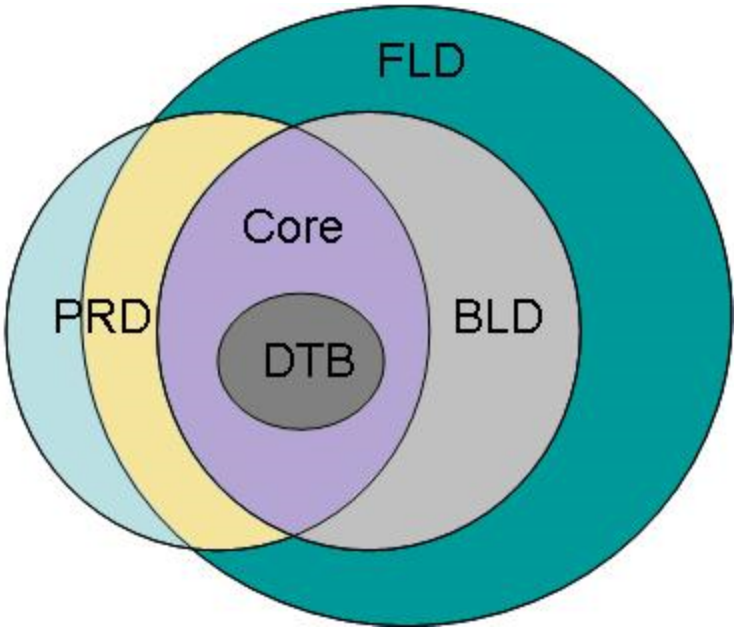
RFI neboli Rule Interchange Format je doporučení konsorcia W3C a slouží jako část infrastruktury sémantického webu. Jedná se o výsledek snahy konsorcia W3C zaměřenou na vývoj standardu pro výměnu dat při zpracovávání informací popsaných soustav0ou pravidel[21]. Pravidlový systém je pak zápis dat pomocí implementace sémantických pravidel. Potřeba zápisu dat pomocí pravidel, které lze využít pro fungování sémantického webu existuje už od konce 90.let, kdy byl tento standard představen.

Vzhledem k existenci současného pravidlového systému a prudkému vývoji jednotlivých technologií, které sice fungovali na jeho bázi, ale například používali pro jejich zápis různé způsoby. Docházelo tak ke vzájemné nekompatibilitě, což byl impuls pro další vývoj RFI jako formátu pro práci s daty při využívání pravidlového systému.

Navzdory svému názvu se ovšem nejedná pouze o formát, resp. standard. Jedná se také o rozšiřitelný framework[22] pro ostatní jazyky využívající pravidlový systém, kterému se říká dialekty. Dialekty obsahují přesné a formální pravidla zápisu pro jednotlivá pravidla. Vztahy mezi různými dialekty lze zobrazit na Vennovým diagramu (viz. obrázek 6).

[21] Vyjádření informace na základě elementární logiky, např.pomocí podmínky If-Then. Používání pravidel je poměrně snadný způsob pro zápis dat. Jedná se o rozklad informací na jednotlivá pravidla, která jsou vyjádřitelná v základní predikátové logice.

[22] Framework je softwarová struktura, která slouží jako podpora při programování a vývoji a organizaci jiných softwarových projektů. Může obsahovat podpůrné programy, knihovny API, podporu pro návrhové vzory nebo doporučené postupy při vývoji.



Obrázek 6: Struktura RFI - Vennův diagram
(Zdroj: http://www.w3.org/TR/rif-ucr/#Structure_of_RIF)

DTB (Datatypes and Built-Ins) – specifikují jednotlivé seznamy typů dat a dalších přídavků, které mají, nebo jsou podporovány RIF dialekty

RIF Core –poskytuje standard pro základní funkcionalitu samotné výměny dat mezi jednotlivými dialekty

RIF FLD (Framework for Logic Dialects) – popisuje jednotlivé mechanismy pro specifikaci nových i současných syntaxí a logických dialektů

RIF BLD (The Basic Logic Dialect) – specifikace FLD schopná realizace Hornových klauzulí[23]. Přidává i další funkce, které nejsou přímo dostupné jako objekty a rámy, IRI, typy dat XML Schema atd.

RIF PRD (Production Rules Dialect) – specifikuje jednotlivé dialekty, aby byli vzájemně schopné si vyměňovat data.

[23] Druh klauzole tvořená disjunkcí literálů ve výrokové logice, která obsahuje nejvýše jeden pozitivní literál. Lze ji tak zapsat jako implikaci a tvořit na jejím základě další výroky. Např. výrok Jsi bud jsi smrtelník, nebo nejsi člověk. Z čehož plyne, že pokud je někdo smrtelník, tak je i člověk.

3.3.7 SWRL

SWRL neboli Semantic Web Rule Language se stejně jako SPARQL používá pro práci s daty ve formátech RDF a dalších založených na OWL. Funguje podobně jako RFI, ale jeho síla tkví v kombinaci OWL a RuleML.

RuleML (Rule Markup Language) – značkovací jazyk vyvinutý pro současné vyjádření 2 způsobů čtení pravidel v XML kvůli dedukci, kvůli zlepšení interpretace dat. Prvním z přístupů je shora dolů (Top-down), který v kontextu zaměření práce znamená rozdělení kódu na jednotlivé malé části, které jsou teprve postupně spojovány do větších celků. Opakem je zdola-nahoru (bottom-up), kdy se nejdříve definují základní elementy, jež jsou následně doplňovány o další funkce, které na nich stavějí.

Implementace RuleML umožnuje pokročilé funkce jako dedukce nových pravidel na základě těch současných. Vznikl na základě iniciativy „Rule Markup Inititive“, která byla zformována kolem snahy vyvinout kanonický[24] webový jazyk pro pravidla, který využívá XML značky způsob transformace z nebo do pravidlového systému.

[24] Kanonický tvar nebo také normální tvor resp. forma objektu označuje tvar, resp. formu ve které může být objekto jednoznačně prezentován.

3.4 Shrnutí

Spojením předcházejících technologií do jednoho soukolí získáme strukturu schopnou analyzovat a procovat s ontologií v rámci webu. URI společně s Unicode tvoří základní kameny, kdy URI jednoznačně identifikuje elementy, se kterými pracujeme a díky znakové sadě je můžeme interpretovat.

XML ačkoliv samo neobsahuje žádné sémantické prvky tak představuje platformu pro jejich ukládání. XML Schema slouží pro definici omezení zápisu strukturovaných dat a díky jmenným prostorům nám umožňuje pracovat s více objekty najednou.

RDF slouží jako datový model pro objekty (zdroje) a vztahy mezi nimi. Také poskytuje základní sémantické prvky, které je možné prezentovat v rámci XML. RDF Schema podobně jako XML Schema představuje slovník, které popisuje schéma vlastností jednotlivých objektů. Tyto základní sémantické prvky jsou následně doplněny OWL, které přidává další slovníky, umožňující komplexnější práci v rámci přiřazování ontologií, jako např. kardinalitu (mohutnost).

Na principu RDF jsou založeny veškeré další technologie pro vytváření sémantického webu. Mezi ty základní patří dotazovací jazyk SPARQL, RIF (Rule Interchange Format) pro vytváření pravidel na sémantickém webu a ontologické struktury RDF Schema a OWL. Ty definují slovníky pro interpretaci sémantiky informací a díky nim je možné odvozovat další informaci aplikováním použitelné logiky.

Velice důležitou součástí dokumentů na webu jsou dále digitální podpisy a šifrování. Ty zajišťují důvěryhodnost a autenticitu jednotlivých dokumentů. Díky digitálnímu podpisu se dají též snadno detekovat případné změny v dokumentu.

Vrchní vrstvy logika a důkaz nejsou prozatím vyřešenou a uzavřenou záležitostí a skupiny pracovníků z W3C na nich stále pracují. Logika na sémantickém webu slouží k automatickému odvozování informací z ontologií a sémantických dat. Díky důkazu by pak mělo být možno přesně určit, zda jsou získané informace pravdivé.

Vrstvou zastřešující všechny ostatní je důvěra. Té bude moci být dosaženo teprve tehdy, až budou vyřešeny všechny ostatní vrstvy a bude potvrzena validita zdroje informací.

V myšlence o sémantickém webu vystupují zásadním způsobem pro práci s ním dvě technologie: XML (eXtensible Markup Language) a RDF (Resource Description Framework). Spojením těchto dvou technologií pak vzniká jeden ze základních jazyků pro zapisování informací na sémantickém webu RDF/XML.

V dnešní době existuje poměrně velké množství alternativ RDF/XML, které jsou na tomto formátu založeny a které ho doplňují a rozšiřují. Mezi nejdůležitější patří například RDFS (RDF Schema), OWL (Web Ontology Language), SKOS (Simple Knowledge Organization System) či Turtle (Terse RDF Triple Language).

Zvláštní postavení mezi technologiemi používanými na sémantickém webu má pak dotazovací jazyk SPARQL, což je jazyk navržený speciálně pro vytváření dotazů na RDF.

4 Praktická část

V praktické části se práce zabývá způsoby uložení dat v podobě strukturované databáze schopné dalšího dotazování a extrakce relevatních výsledků prostřednctvím dotazů. Pro praktické potřeby a třídění dat uložených v rámci dokumentů práce identifikuje nástroje pro jejich transformaci z jednoduchých dat do strukturované podoby a nalezení rozdílů mezi těmito nástroji.

4.1 Linked Data

Internet byl vždy způsobem, kde se přes počítač mohl uživatel připojit k různým zdrojům informací. Na začátku to byly různé dokumenty na webových stránkách. Tyto dokumenty ale postupně začali v sobě propojovat další informace z jiných zdrojů. To nám dávalo možnost procházet přes odkazy na různé zdroje z jednoho místa. Tak se jeden dokument začal skládat z několika zdrojů, kdy se například u stránce hudební kapely, krom původního textu, objevuje i obrázky alb, nebo samotné ukázky jejich tvorby, ačkoliv tyto informace mohou být nahrány na jiné adrese a v rámci stránky figurovat pouze jako zobrazené odkazy, podobně jako například funguje vložení obrázku do HTML formou odkazu na jeho umístění[25]. Toto je možné díky unikátním identifikátorům, pomocí kterých můžeme tyto data kombinovat mezi sebou. Dalším stupněm vývoje bylo připojení těchto dat k aplikační službě (API), kdy už se uživatel stal aktivním prvkem. API mu umožňuje procházet a měnit samotná data skrz různé webové aplikace. Místo pouhého prohlížení informací mohl tak začít tvořit nové, například na sociálních sítí, blozích apod. Aby ovšem tato data byla přístupná, musí být poskytována v jednotné formě kvůli kompatibilitě, tato forma se nazývá Linked Data.

Toto vše je umožněno díky několika jednoduchým pravidlům. Komunikačním standardem na webu je http, neboli Hypertext transfer protocol, pomocí nějž dochází k získávání samotných dat. Způsob jakým je kontrolováno jak se data a média zobrazí je kontrolován jazykem HTML (Hypertext Markup Language). Způsob jakým nalezneme

[25] <src href="adresa_obrázku">

jednotlivé webové stránky pomocí názvu je realizován pomocí URL (Uniform Resource Locator).

Právě tyto technologie umožňují počítač procházet data novým způsobem. Jednodušším a rychlejším způsobem. Tim Berners-Lee, který stál společně se svou organizací W3C u původu Linked Data představil 3 základní pravidla pro Linked Data. Toto se stalo na konference TED v roce 2009.

- Kterýkoliv objekt na webu má mít přiřazený identifikátor URI – jedná se o unikátní perzistentní identifikátor, který se tak nebude měnit a bude neustále k dispozici, jedná se o adresu na místo, kde se daná data nachází
- Tyto identifikátory vrací zpátky požadované informace ve standardním formátu
- Tyto informace budou obsahovat explicitní vazby na jiné objekty odkazované pomocí URI – to je okamžik, kdy se obyčejná data transformují na linked Data (provázaná data)

Tyto data jsou tak poskytována ve formě trojicí, která zabezpečuje kompatibilitu s formátem RDF. Pomocí trojicí je možné definovat jakoukoliv vlastnost objektu. Příkladem stránky kde se již tento formát využívá je www.bbc.co.uk/nature. Spíše než poskytování předem definovaných statických textů využívá data, která převážně pocházejí z linked data skladiště na jiných stránkách. Tyto data vyhledává na základě udaných trojic. Aplikace dostanou dokonce práva tvořit i nové trojice na základě dat. Udané stránky bbc/nature jsou koncipovány do dvou částí, kdy se v horním uživatel může dočíst nejnovější články o zvířatech, zatímco v dolní části se automaticky vyplňují informace o zvířeti, které v daném článku vystupuje a o dalších článcích o něm. V příkladu vrány na stránkách bbc nature tak dochází ke tvorbě např. trojice, kdy vrána patří do rodu vrána obecná. Pokud klikneme na odkaz vrána obecná, tak také vytvoří trojici, kdy vrána obecná obsahuje jednotlivé vrány. V případě, že se v novinkách objeví nový článek o vranách, tak o tom jednotlivé aplikace jsou informovány, na základě trojicí tak zjistí vlastnosti vrány a sami vytvoří stránku, kterou naplní informacemi a odkazy na články o vranách. Je to dáno aplikacemi, které dokážou mezi sebou komunikovat prostřednictvím trojic.

Typickým příkladem linked Data jsou projekty jako Dbpedia či Freebase. Jedná se tak o databázi strukturovaných dat, která jsou volně přístupná.

Uchovávání informacích v takovýchto knihovnách poskytuje hlavní 3 výhody:

- Lidé snadněji nachází data na webu
- Více aplikací využívající metadata
- Kategorizace jednotlivých pojmů na základě vlastností

Zde je zejména důležitý třetí bod. Pokud má každý objekt dané vlastnosti, tak je můžeme na jejich základě i kategorizovat v mnohem účinnější formě. Umožňuje to schopnost pamatovat si již dříve vytvořené trojice. Toto je zvláště důležité, pokud se setkají totožné pojmy s různými vlastnostmi. Pokud tedy máme pojem „vrána“, ke kterému jsou přiřazena metadata, která značí, že jejím autorem je Allan Edgar Poe, či že patří mezi romány, tak počítač je schopen tento pojem odlišit od vrány ve smyslu zvířete. Toto se děje na základě kontextu, kdy počítač zjišťuje, k čemu se vztahují okolní pojmy a porovnává je se svoji databází, dokud nenajde výsledek s odpovídajícími vlastnostmi.

Tyto explicitní vztahy mezi jednotlivými objekty vyžaduje určité dohody ve způsobu zápisu ontologie, kterou jednotlivé slovníky a datové strukturu poskytují. Viz. kapitola Přílohy.

Využitím těchto knihoven trojic při vyhledávání dojde ke zlepšení relevantnosti výsledků vyhledávání. Při vyhledávání se tak informace z různých míst, stránek a adres seskupují na jednom místě, v jedné knihovně. Tyto data se navíc transformují do standardu, který je dobře strojově zpracovatelný.

Na základě těchto strojově čitelných informací lze tak snadno vytvářet aplikace, která umí využívat několik zdrojů dat najednou. Například udělat video aplikaci, která krom zobrazování videa popisující vránu, také dokáže zobrazovat text o vranách pocházející z úplně jiného zdroje. K němu může být přiřazena nahrávka, opět z jiného zdroje, z nichž daný text jako přepis vychází atd. Tyto aplikace se nazývají mash-up.

4.1.1 Ontologie

Ontologie původně znamená filosofickou disciplínu, která se zabývá jsoucnem a bytím jako takovým. Umožňuje definování pojmů na základě přiřazení jejich vlastností.

Pro vývojáře sémantického webu tak ontologie znamená soubor, který definuje vztahy mezi jednotlivými pojmy. Jinak řečeno ontologie představuje možnost tvorby metadat, neboli informací o informacích. Obsahuje v sobě taxonomii a odvozovací pravidla. Taxonomie nám umožní definování jednotlivých tříd objektů a vztahů mezi nimi. Právě tyto třídy obohacené o své vlastnosti představují velmi mocný nástroj. Přiřadíme-li třídám určité vlastnosti a zároveň povolíme podtřídám tyto vlastnosti dědit, můžeme mezi objekty vyjádřit obrovské množství vztahů. Navíc odvozovací pravidla nám umožní poté vyvodit další vztahy, např. Auto řídí člověk a Petr řídí auto. Z toho lze odvodit, že Petr je člověk. Účelem ontologie je tak podpora porozumění mezi lidmi, podpora komunikace mezi systémy a podpora návrhu znalostně orientovaných systému.

Ontologie tak velkou mírou zlepšuje a rozšiřuje fungování webu. I když počítač nikdy skutečně neporozumí významu slova, dokáže si k němu přiřadit vlastnosti, díky nimž dokáže pracovat podstatně efektivněji. Výsledky vyhledávání se zlepšují. Vyhledávací program tak dokáže vyhledat pouze ty stránky, které odpovídají významu hledaného výrazu. Zjišťuje, zda opravdu odpovídají hledanému výrazu. Zajímá se tak o jeho význam a klíčové slovo tak přestává být pouze soubor znaků, které porovnává s ostatními dokumenty na webu a hledá shodu. Klíčové slovo bývá často nejednoznačné a může reflektovat více pojmů.

4.1.1.1 Typy ontologií

terminologické - pokročilejší tezaury[26]. Používané jsou v knihovnictví a oborech zaměřených převážně na textové informace.

informační - rozvinutí databázových konceptuálních schémat. Zajišťují abstrakci a vyšší kontrolu integrity

znalostní - reprezentace znalostí v rámci umělé inteligence. Objekty a relace mezi objekty jsou důsledně definovány pomocí formálního jazyka.

[26] Tezaurus – slovník deskriptorů, slovních výrazů sloužících k popisu obsahu dokumentů a jejich následném vyhledávání; spolu s nedeskriptory (soubor výrazů, které deskriptor nahrazuje) tvoří slovníkovou jednotku tezauru; jsou v něm určeny vztahy nadřazenosti a podřazenosti, termíny synonymní, antonymní apod.

4.1.2 Příklady Linked Data

Pro definici pojmů a vztahů se v ontologii (rozumějme v rámci zaměření této práce) používají tzv. slovníky. Ty na sémantickém webu pojmy a vztahy používané k popisu a reprezentaci oblasti zájmu. Slovníky jsou používány ke klasifikaci pojmů, které mohou být použity ve speciální aplikaci, charakterizovat možné vztahy a vymezit případná omezení používání těchto pojmů. V praxi mohou být slovníky velmi komplexní a mohou obsahovat tisíce pojmů, nebo jsou velmi jednoduché a popisují například jen jeden či dva pojmy. Slovníky jsou základními stavebními kameny řídících technik na sémantickém webu. Slovníky jsou na sémantickém webu důležité především kvůli pomoci při propojení dat, například pokud by se mohly vyskytovat nějaké nejasnosti v pojmech používaných v různých datových souborech nebo kdyby nějaké další poznatky mohly vést k objevu nových vztahů.

Jako příklady schémat ontologií jsou využity následující příklady linked data.

Dbpedia – http://mappings.dbpedia.org/server/ontology/classes/

Freebase - http://schemas.freebaseapps.com/

4.1.2.1 Dbpedia

Dbpedie je obdoba wikipedie zapsaná v RDF standardu přes trojice. Jedná se o knihovny podobné databázím. Momentálně obsahuje 3,77 milionu elementů s více než 400miliony fakty pouze v angličtině. Celkově je ovšem dostupná ve 111-ti jazycích což dohromady tvoří popis 20,8 milionu objektů.

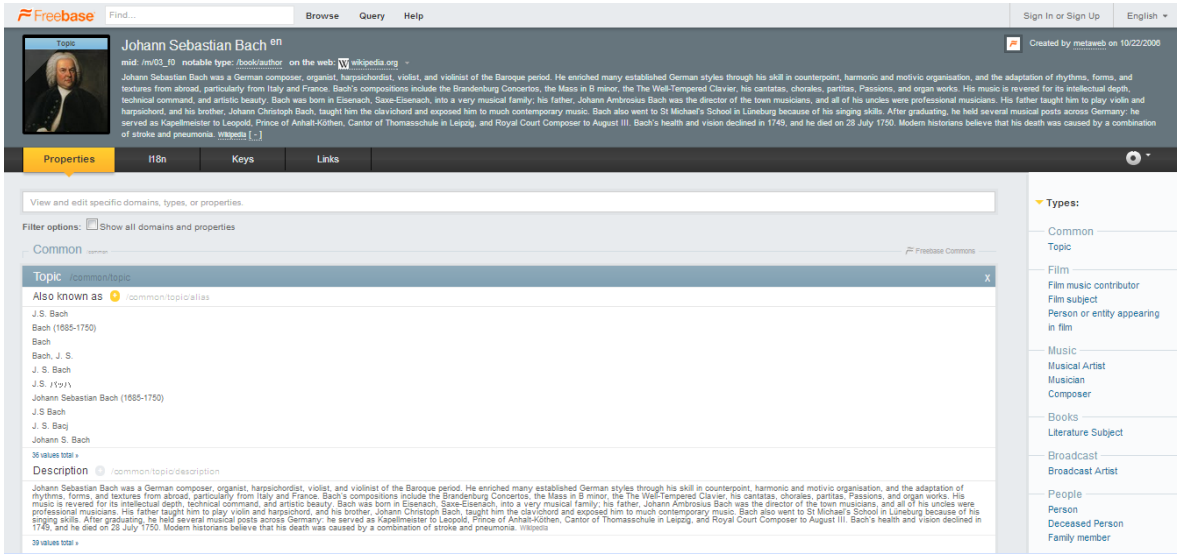
Jako příklad byl vybrán zápis objektu popisující německého skladatele Johanna Sebastiana Bacha. Tato databáze v RDF standardu vypadá jako seznam atributů a k nim přiřazené hodnoty (viz. obrázek 8).

Obrázek 7: Dbpedia – popis J.S.Bacha v prostředí Dbpedia (Zdroj: http://dbpedia.org/page/Johann_Sebastian_Bach)

4.1.2.2 Freebase

Freebase je podobně jako Dbpedie rozsáhlou znalostní databází skládající se z metadat, které byli seskládány členy její komunity. Skládá se z dat, která pocházejí z mnoha zdrojů. Jejím cílem je vytvoření globální zdroje informací, které budou čitelné strojem i člověkem. Stojí za ní společnost Metaweb a funguje veřejně již od května 2007. Momentálně obsahuje přes 23 miliónů entit.

Na rozdíl od klasických znalostních databází Freebase neukládá data ve formě tabulek, ale grafu. Data definuje jako uzly grafu a hrany, které je spojují a definují tak vztahy mezi nimi. Díky tomu celá koncepce ukládání dat postrádá hierarchické rozřazení, může Freebase definovat mnohem komplexnější vztahy.



Obrázek 8: FreeBase - J.S.Bach
(Zdroj: www.freebase.com)

4.1.3 Extrakce dat

Jako příklad pro extrakci dat byl vybrán jako zdroj databáze Dbpedia a jako nástroj dotazovací jazyk SPARQL v prostředí Virtuoso SPARQL Generator[27]. V následující příkladu je databáze dotazována na popis slavného skladatele Johanna Sebastiana Bacha v anglickém jazyku. Výsledek je poté zobrazen na obrázku 9.

```
select ?abstract where {{
<http://dbpedia.org/resource/Johann_Sebastian_Bach>
<http://dbpedia.org/ontology/abstract> ?abstract .
FILTER langMatches( lang(?abstract), 'en') }}
```

SPARQL results:

abstract
"Johann Sebastian Bach (31 March 1685 – 28 July 1750) was a German composer, organist, harpsichordist, violist, and violinist of the Baroque Period. He enriched many established German styles through his skill in counterpoint, harmonic and motivic organisation, and the adaptation of rhythms, forms, and textures from abroad, particularly from Italy and France. Bach wrote much music, which was revered for its intellectual depth, technical command, and artistic beauty. Many of his works are still known today, such as the Brandenburg Concertos, the Mass in B minor, the Well-Tempered Clavier, and his cantatas, chorales, partitas, passions, and organ works. Bach was born in Eisenach, Saxe-Eisenach into a very musical family; his father, Johann Ambrosius Bach was the director of the town's musicians, and all of his uncles were professional musicians. His father taught him to play violin and harpsichord, and his brother, Johann Christoph Bach taught him the clavichord, and exposed him to much contemporary music. Bach also sang, and he went to the St Michael's School in Lüneburg, because of his skill in voice. After graduating, he held several musical posts across Germany; he served as Kapellmeister (director of music) to Leopold, Prince of Anhalt-Köthen, Cantor of Thomasschule in Leipzig, and Royal Court Composer to August III. Bach's health and vision declined in 1749, and he died on 28 July 1750. Modern historians believe that his death was caused by a combination of stroke and pneumonia. Bach's abilities as an organist were highly respected throughout Europe during his lifetime, although he was not widely recognised as a great composer until a revival of interest and performances of his music in the first half of the 19th century. He is now generally regarded as one of the main composers of the Baroque period, and as one of the greatest composers of all time."@en

Obrázek 9: Výsledek dotazu J.S.Bach
(Zdroj: http://dbpedia.org/snorql/)

[27] http://dbpedia.org/sparql

Odpověď dotazu lze v závislosti na uživatelově požadavku interpretovat ve více jazycích. V tomto případě byl vybrán standard RDF/XML.

```
<rdf:RDF xmlns:res="http://www.w3.org/2005/sparql-results#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
<rdf:Description rdf:nodeID="rset">
<rdf:type rdf:resource="http://www.w3.org/2005/sparql-results#ResultSet" />
   <res:resultVariable>abstract</res:resultVariable>
   <res:solution rdf:nodeID="r0">
    <res:binding
rdf:nodeID="r0c0"><res:variable>abstract</res:variable><res:value
xml:lang="en">Johann Sebastian Bach (31 March 1685- 28 July 1750) was a
German composer, organist, harpsichordist, violist, and violinist of the Baroque
Period. He enriched many established German styles through his skill in
counterpoint, harmonic and motivic organisation, and the adaptation of rhythms,
forms, and textures from abroad, particularly from Italy and France. Bach wrote
much music, which was revered for its intellectual depth, technical command, and
artistic beauty. Many of his works are still known today, such as the Brandenburg
Concertos, the Mass in B minor, the Well-Tempered Clavier, and his cantatas,
chorales, partitas, passions, and organ works. Bach was born in Eisenach, Saxe-
Eisenach into a very musical family; his father, Johann Ambrosius Bach was the
director of the town´s musicians, and all of his uncles were professional musicians.
His father taught him to play violin and harpsichord, and his brother, Johann
Christoph Bach taught him the clavichord, and exposed him to much contemporary
music. Bach also sang, and he went to the St Michael´s School in LĂĽneburg,
because of his skill in voice. After graduating, he held several musical posts across
Germany; he served as Kapellmeister (director of music) to Leopold, Prince of
Anhalt-Kothen, Cantor of Thomasschule in Leipzig, and Royal Court Composer to
August III. Bach´s health and vision declined in 1749, and he died on 28 July 1750.
Modern historians believe that his death was caused by a combination of stroke
and pneumonia. Bach´s abilities as an organist were highly respected throughout
Europe during his lifetime, although he was not widely recognised as a great
composer until a revival of interest and performances of his music in the first half
of the 19th century. He is now generally regarded as one of the main composers of
the Baroque period, and as one of the greatest composers of all
time.</res:value></res:binding>
   </res:solution>
  </rdf:Description>
</rdf:RDF>
```

Výsledek byl srovnán se zdrojovými údaji (viz. obrázek 7) a na základě jejich shody byla potvrzena jeho správnost.

4.1.4 Shrnutí

Linked data představují model a základní pravidla pro distribuci strukturovaných dat, tak aby bylo dosaženo jejich vzájemné kompatibility a usnadněna práce s nimi.

Představuje další směr vývoje způsobu uložení dat, které umožňují způsob práce podobným relačním databázím a díky nim se již nejedná o seznam dokumentů, kde každý z nich obsahuje popis jednoho objektu, ale o strukturované data, která lze libovolně dotazovat a na základě výsledků teprve poskytovat libovolné informace v uživateli požadované podobě. Jejich základní předpoklad je schopnost odkazování dat mezi sebou a jejich transformace do jedné propojené struktury.

V rámci praktické části byl úspěšně vytvořen dotaz na databázi Dbpedia, který umožňoval extrakci vybraných dat bezu nutnosti použití přímého odkazu na požadované místo uložení dat.

4.2 Transformace na strukturované data

Dnes je webový obsah prezentován v (X)HTML spíše jako „plakát“, může se sice jednat o plakát interaktivní umožňující třeba zápis. Stále je ale stroj naprosté většině obsahu nerozumí. Pokud by se toto změnilo, tak by i pro uživatele bylo mnohem snažší pracovat s obsahem. Jedná se tak zejména o spolupráci prohlížeče s dalšími aplikacemi, resp. předat si mezi sebou data z obsahu webové stránky v požadované podobě. Typicky se tak může dít v případě, že uživatel na webové stránce potvrdí svoji účast na události, kdy se tato informace zaznamená diářem, který se sám následně aktualizuje, do seznamu kontaktů se uloží kontakt na hostitele a do GPS se automaticky uloží poloha místa konání události.

Při sémantickém vyhledávání či analýze dat se ve většině případů používají jako zdroje velké znalostní databáze RDF trojic zapsané pomocí standardů RDF/XML či Turtle. Ačkoliv existují přístupy pro začlenění těchto nástrojů přímo do XHTML dokumentu, tak většinou se jedná o poměrně náročná řešení, kdy jsou data povětšinou implementovány ve formě bloků ohraničených např. tagem <sparql> či <rdf> a jsou tak vhodná spíše pro tvorbu znalostních bází např. ve formátu XML, než pro přímou práci v rámci dokumentu. Z toho plyne jejich špatné využití pro transformaci stávajících nestrukturovaných dat na strukturované přímo uvnitř XHTML.

4.2.1 Nástroje

Pro úpravu dat v rámci webových dokumentů se používají nástroje pracující na principu přidávání dalších XHTML značek definující sémantický výraz element, ke kterému jsou přiřazeny. Existují 3 hlavní způsoby: RDFa, Mikrodata a Mikroformáty. Fungují na principu obohacení jednotlivých částí dokumentu o metadata formou tagů (značek), které jim dodatečně přidávají význam. Tyto způsoby nijak neovlivňují zobrazení dokumentu v rámci prohlížeče a jejich využití je nepovinné, vyžaduje vysokou zodpovědnost ze strany tvůrce obsahu, aby dané značky skutečně využíval.

Jako příklad implementace bude u každého nástroje zobrazen převod následujícího textu na strukturovaná data. Pro názornost se jedná o popis autorovi osoby formou vizitky.

Výchozí data pro implementaci:

Jméno: David Macháček
Domácí stránka: www.xmacd108.php5.cz
Telefon: 776 123 456
Email: davido.machacek@gmail.com
Adresa
Ulice: Lessnerova
Město: Praha
Stát: Česká Republika
Zaměstnání
Pozice - Mobile Content Assistant
Společnost - Telefónica O2

4.2.1.1 RDFa

Jako jeden ze způsobů zápisu datového formátu RDF se využívá RDFa (Resource Definition Framework in attributes). Jedná se tak o technologii pro přenos strukturovanou informaci skrz webovou stránku. Slouží tak pro implementaci RDF do (X)HTML. RDFa tak umožňuje zápis vlastností objektů uvnitř webového dokumentu. Používá tak atributy jak z (X)HTML (src, href a další), tak k nim i přidává vlastní vlastnosti (about, property, resource..). Její nejdůležitější vlastnost je nutnost odkazu na platný jmenný prostor, který zároveň i definuje typ objektu, který je popisován. To se děje pomocí RDF schéma.

Příklad implementace:

```
<div xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns="http://www.w3.org/1999/xhtml"
    xmlns:rdfa="http://www.w3.org/ns/rdfa#"
    xmlns:md="http://www.w3.org/ns/md#"
    xmlns:schema="http://schema.org/"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"class="rdf2rdfa">
<div class="description" about="http://www.xmacd108.php5.cz.html">
    <div rel="md:item">
      <div xmlns="" typeof="rdf:List">
        <div rel="rdf:first">
        <div xmlns="http://www.w3.org/1999/xhtml" class="description"
        typeof="schema:Person">
            <div rel="schema:address">
              <div class="description" typeof="schema:PostalAddress">
```

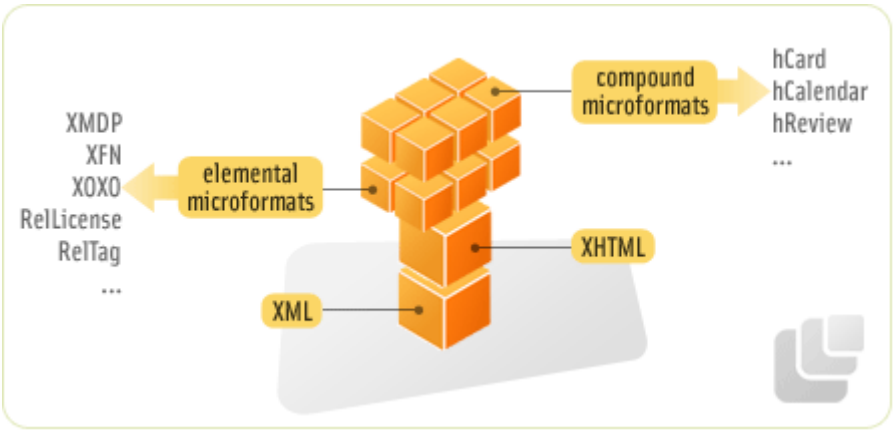
```
            <div property="schema:streetAddress" content="Lessnerova"/>
              <div property="schema:addressLocality" content="Praha"/>
              <div property="schema:addressCountry" content="Česká
              Republika"/>
            </div>
          </div>
          <div property="schema:affiliation" content="Telefónica O2"/>
              <div property="schema:email"
              content="davido.machacek@gmail.com"/>
          <div property="schema:jobTitle" content="Mobile Content
          Assistant"/>
          <div property="schema:name" content="David Macháček"/>
          <div property="schema:telephone" content="776 123 456"/>
        <div rel="schema:url"
        resource="http://www.xmacd108.php5.cz/www.xmacd108.php5.cz"/>
          </div>
        </div>
        <div rel="rdf:rest" resource="[rdf:nil]"/>
      </div>
    </div>
    <div rel="rdfa:usesVocabulary" resource="http://schema.org/"/>
  </div>
</div>
```

Jak je vidět, jedná se o velice robustní formát. RDFod počátku nebylo příliš přívětivé k vývojářům. Jeho cílem je dostat RDF na web v plné síle. Ne vždy je ovšem vyžadovaná vysoká schopnost expresivity a často se tedy pak před ním mikroformáty. Ty jsou mnohem kompaktnější a snadněji definovatelné.

4.2.1.2 Mikroformáty

Proces sémantizace webu není žádná náhlá revoluce v evoluci webu. Jedná se o pomalý proces postupného využívání vícero technologií, např. XML, které vzájemnou symbiózou vytváření komplexní myšlenku sémantického webu. Jeho nevýhoda je tak nutnost donutit vývojáře a tvůrce samotných stránek k jejich využívání. Většina webu je však stále psána v jednoduchém (X)HTML, které neobsahuje sémantické prvky potřebné ke strojově čitelnému vyjádření významu jednotlivých elementů. Bylo tedy třeba přidat způsoby jakým jednoduše přidat sémantické prvky do (X)HTML, tuto myšlenku se podařilo naplnit ve standardu, který dnes nazýváme mikroformáty.

Mikroformáty tak díky svoji schopnosti přidávat informace o informacích můžeme považovat za metainformace. Jejich použitím lze dosáhnout lepších výsledků při vyhledávání, a práce s daty přes prohlížeč. Tam kde jiné formy sémantizace fungují jako zdroj pro budování následné prezentace v (X)HTML, tam mikroformáty naopak zastávají pozici nadstavby již hotové prezentace o metadata. Toto je zachyceno na následujícím schématu.



Obrázek 10: Struktura mikroformátů
(Zdroj: https://encrypted-tbn2.gstatic.com/images?q=tbn:ANd9GcQhyt-x9No6QWKHzYsBlBbn6cd_enDW-GY-qbx20OdKkXBcYs-9Jg)

Průkopník, který stál za vznikem mikroformátů, byl Tantek Çelik, který v letech 1997-2004 pracoval v Microsoftu mimo jiné jako zástupce ve W3C. Snažil se velkou měrou prosadit jednoduchý formát XFN (XHTML Friends Network), který připravil společně s Erickem Meyerem a Matthew Mullunwegem, pozdějšími spoluzakladateli Global Multimedia Protocols Group, zabývajícími se experimentální matematikou. Fungoval na jednoduchém principu popisování lidských vztahů pomocí odkazů. Umožnuje tak pomocí přidání malého atributu definovat svůj vztah k jiné osobě. Například, kdy označím danou osobu jako svého přítele:

Honza Novak

Prosazení XFN se mu ovšem nepodařilo, což mimo jiné zapříčinilo jeho odchod z W3C. K tomuto formátu se následně připojili další, jako XMDP, XOXO, RelLicense, RelTag a společně dnes tvoří mikroformáty. Sdružují se do organizace microformats.com, což je naprosto otevření iniciativa, narozdíl od do jisté míry „elitářského“ W3C.

V dnešní době již jsou mikroformáty široce rozšířené a používají je i velké společnosti jako Facebook, Twitter, Wikipedia, Microsoft či Google. Jejich široké rozšíření dokládá i vyvinutí standardu GRDDL (Gleaning Resource Descriptions from Dialects of Languages), které popisuje způsob extrakce RDF dat z XML a zvláště (X)HTML dokumentů, včetně mikroformátů.

Mikroformáty či µf, jak se někdy zkracují, se zapisují pomocí přidaných atributů do tagů uvnitř stránky. Jako příklad uvádím dva nejpoužívanější atributů, rel a class.

4.2.1.2.1 Rel

Popisuje vztah mezi cílovou a zdrojovou stránkou, může obsahovat více hodnot oddělených mezerou.

Home – označuje domovskou stránku či může upřesňovat strukturu webu ukazováním na hlavní stránku, např. „*Hudební kapela“*

Me – odkazuje na jinou stránky obsahující informace o té stejné osobě, využití např. ze sociálních sítí, např. „*Můj profil“*

Tag – používáno pro jakési „označkování“ stránky pro usnadnění kategorizace a vyhledávání na webu. Slouží pro označení stránky záložkou, která bude na danou stránku odkazovat, např. „*Nejrozšířenější český vyhledávač“*

Licence – odkazuje na licenční ujednání pod jakým je publikovaný obsah stránky, např. „*//creativecommons.org/licenses/by/3.0/us/“ rel=”licence“ >Licenční ujednání“*

4.2.1.2.2 Class

Tento atribut umožňuje tvorbu vlastních tříd ve struktuře metadat. Pojmenování těchto tříd závisí na programátorovi. Jelikož tak každý vývojář mohl pojmenovávat stejná data jiným způsobem, tak se objevily problémy se vzájemnou kompatibilitou a z toho

důvodu byly vyvinuty standarda pro základní typy tříd jako hCard (osoba), hNews (zpravodajství), hProduct (produkt), hReview (hodnocení) či hResume (životopisy) a další. Také je ovšem zdroje veliké nevýhody mkroformátů, jelikož tento tag koliduje s využití CSS stylů a komplikuje jejich využití.

hCard

hCard je standard pro definování struktury dat při popisu elektronických vizitek. Mikroformáty jej využívají jako schéma pro popis vlastností jednotlivých osob. Využívají pro to třídu vCard. Funguje jako formát pro výměnu osobních dat, především jako objekt s vlastnostmi jako jméno a příjmení, adresy, telefonní čísla a další. Jsou ukládány ve formátu .vcf a v současné době jsou již k dispozici hojně na mobilních zařízení a na emailu ve formě vizitek. Díky formátu hCard tak lze obsažená data konvertovat do vCard a následně využít v programech jako Microsoft Outlook či Address Book pro ulehčení tvorby kontaktů.

Příklad implementace:

```
<div class="vcard">
Jméno: <span class="fn">David Macháček</span>
Domácí stránka: <a href="www.xmacd108.php5.cz"
class="url">www.xmacd108.php5.cz</a>
Telefon: <span class="tel">776 123 456</span>
Email: <span class="email">davido.machacek@gmail.com</span>
Adresa:
<span class="adr">
        <span class="street-address">Lessnerova</span>
        <span class="locality">Praha</span>,
        <span class="country">Čechy</span>
</span>
Zaměstnání:
<span class="title">Mobile Content Assistant</span>
<span class="org">Telefónica O2</span>.
</div>
```

Další příklady podobných standardů jsou např. hCalendar, hReview a další.

Mikroformáty tak tvoří zajímavý, snadno implementovaný způsob, jak přiřadit sémantický význam jednotlivým elementům. Jejich účelem nikdy nebylo schopnost vysoké expresivnosti, ale snadný a praktický přístup k sémantizaci.

4.2.1.3 Mikrodata

V této práci již byli představeny 2 základní způsoby jak doplnit obsah webového dokumentu o sémantické značky. Jedná se tak zejména o využití komplexního RDFa či případně nadstavby ve formě jednoduchých mikroformátů využívající atributy *class* a *rel* v (X)HTML. Obě technologie mají své kladné i záporné stránky. Zejména u RDFa, které je náročné svoji syntaxí, či u mikroformátů díky jejich omezené schopnosti definovat náročnější vlastnosti. Informace získané při vývoji těchto dvou technologií vedli ke vzniku jejich třetí alternativy, tedy služby vhodné k obohacování dat o sémantiku prostředky, které nejsou jádrem (X)HTML, tato technologie se nazývá mikrodata.

Mikrodata fungují na principu slovníků. Představme si seznam všech elementů v rámci HTML 5 jako seznam slovníků. Tyto slovníky jistě budou obsahovat elementy pro reprezentaci odstavce či názvu článku, ale již ne elementy pro osobu, událost apod. Pokud tak chceme vytvářet na webu událost se seznamem osob, které se ji hodlají účastnit, tak již budu tyto možnosti potřebovat. Jednou z alternativ je tak tvorba nových, vlastních slovníků, které jsou obohaceny o nové, námi požadované elementy. Toto nám mikrodata umožňují.

Pracují na principu dvojic název/hodnota, kdy jsou následně jednotlivé elementy obohaceny o další pojmenované vlastnosti. Například tak osoba může být obohacena o vlastnosti jako jména, adresa a další. Mikrodata využívají hierarchickou strukturu (X)HTML dokumentu (např. element <HTML> obsahuje <head> a <body>). Zanořováním tagů dovnitř objektového modelu dokumentu tak vyjadřují, že „všechny vlastnosti uvnitř *tohoto* elementu jsou vzaty z *tohoto* slovníku.“ Díky tomu můžeme v rámci jednoho dokumentu využívat více slovníků. Dokonce vkládání slovníků dovnitř již jiných.

Definování vlastní slovníků je velice jednoduché. Pro každý slovník je nejdříve třeba uvést jeho jmenný prostor ve formě URL. Ta může vést na fungující web, ale není to striktně požadováno. Pokud tak chci vytvořit slovník popisující osobu a vlastním stránku www.mojestranka.cz tak jmenný prostor pro nový slovník může být definován jako www.mojestranka.cz/osoba. Tím zajistím jeho globální unikátnost. Po definování jmenného prostoru již mohu přejít k samotné syntaxi. Pro značení kódu se používají následující atributy:

Itemscope – slouží pro definování nového objektu

Itemtype – odkazuje na schéma určující typ objektu a slovník jeho vlastností

Itemprop – funguje jako popis atributu

Zkombinování těchto atributů se dosahuje popisu vlastností a podstaty jednotlivých elementů uvnitř dokumentů.

Příklad implementace:

```
<div itemscope itemtype="http://schema.org/Person">
Jméno: <span itemprop="name">David Macháček</span>
Domácí stránka: <a href="www.xmacd108.php5.cz"
itemprop="url">www.xmacd108.php5.cz</a>
Telefon: <span itemprop="telephone">776 123 456</span>
Email: <span itemprop="email">davido.machacek@gmail.com</span>
Adresa:
<span itemscope  itemprop="address"
itemtype="http://schema.org/PostalAddress">
        <span itemprop="streetAddress">Lessnerova</span>
        <span itemprop="addressLocality">Praha</span>
        <span itemprop="addressCountry">Česká Republika</span>
</span>
Zaměstnání:
        <span itemprop="jobTitle">Mobile Content Assistant</span>
        <span itemprop="affiliation">Telefónica O2</span>.
</div>
```

Mikrodata tak tvoří zajímavou alternativu k RDF a mikroformátům. Jejich implementace nikterak nemění vzhled stránky a taktéž jejich syntaxe je snadná.

4.2.1.3.1 Schema.org

Jedná se o společný projekt společností Google, Bing! a Yahoo. Jejím účelem je poskytnutí schémat, které umožní porozumění obsahu hlavními hráči na poli vyhledávačů. Jeho účelem je definování různých typů entit na stránce. Například pokud vytvořím objekt s vlastnosti jako jméno a země původu, prohlížeč má velmi omezené možnosti identifikace jestli se jedná o osobu, nebo třeba výrobek. Z toho důvodu byla vyvinuta databáze objektů,

které mají předdefinované vlastnosti. Jejím použitím jasně definuje zařazení obsahu a vlastností objektu. Umožníme tak jejich porovnávání a usnadníme práci s nimi.

Obsahuje v sobě velké množství široce použitelných typů objektů jako osoba, událost, organizace, místo, hodnocení, a tak dále. Společně s nimi ovšem také nabízí i více specializované typy objektů jako TV seriál, socha či dokonce úzce oborové jako medicína. Celý seznam lze najít na http://schema.org/docs/full.html.

Thing > Event

An event happening at a certain time at a certain location.

Property	Expected Type	Description
Properties from Thing		
additionalType	URL	An additional type for the item, typically used for adding more specific types from external vocabularies in microdata syntax. This is a relationship between something and a class that the thing is in. In RDFa syntax, it is better to use the native RDFa syntax - the 'typeof' attribute - for multiple types. Schema.org tools may have only weaker understanding of extra types, in particular those defined externally.
description	Text	A short description of the item.
image	URL	URL of an image of the item.
name	Text	The name of the item.
url	URL	URL of the item.
Properties from Event		
attendee	Organization or Person	A person or organization attending the event.
attendees	Organization or Person	A person attending the event (legacy spelling; see singular form, attendee).
duration	Duration	The duration of the item (movie, audio recording, event, etc.) in ISO 8601 date format.
endDate	Date	The end date and time of the event (in ISO 8601 date format).
location	Place or PostalAddress	The location of the event or organization.
offers	Offer	An offer to sell this item—for example, an offer to sell a product, the DVD of a movie, or tickets to an event.
performer	Organization or Person	A performer at the event—for example, a presenter, musician, musical group or actor.
performers	Organization or Person	The main performer or performers of the event—for example, a presenter, musician, or actor (legacy spelling; see singular form, performer).
startDate	Date	The start date and time of the event (in ISO 8601 date format).
subEvent	Event	An Event that is part of this event. For example, a conference event includes many presentations, each are a subEvent of the conference.
subEvents	Event	Events that are a part of this event. For example, a conference event includes many presentations, each are subEvents of the conference (legacy spelling; see singular form, subEvent).
superEvent	Event	An event that this event is a part of. For example, a collection of individual music performances might each have a music festival as their superEvent.

Obrázek 11: Schéma objektu typu událost v schema.org
(Zdroj: www.schema.org)

Podpora velkých společností jako Google, Yahoo a Bing! je velmi důležité. Tyto společnosti jsou vedoucí subjekty na trhu hypertextových vyhledávačů a jejich zaměření na mikrodata velkou měrou určuje další zaměření celého trhu. Vývojáři jsou tak nuceni využívat mikrodata, aby tak maximálně mohli využívat potenciál těchto vyhledávačů a dostali tak své produkty jejich prostřednictvím mezi zákazníky.

Pro doplňující informace je přiložen výtah z vyjádření jednotlivých společností k důvodu zaměření na mikrodata z 2.června 2011, kdy bylo schema.org oficiálně spuštěno:

Google

„V minulosti jsme podporovali 3 různé standardy pro značení strukturovaných dat: mikrodata, mikroformáty a RDFa. Rozhodli jsme se zaměřit pouze na jeden formát prostřednictvím schema.org, abychom mohli vývojářům nabídnout jednodušší nástroje a zajistili kompatibilitu napříč vyhledávači. Pro každý standard existují argumenty, které jej vyzdvihují nad ostatní, ale právě mikrodata nabízejí optimální poměr mezi expresivitou RDFa a jednoduchostí mikroformátů a proto jsme se rozhodli zaměřit na tento formát.“[28]

Yahoo!

„Dnešní oznámení nabízí obrovskou příležitost pro další růst. Kromě konsolidace stávajících schémat pro slovníky, které jsme již podporovali, tak to znamená podporu nových schémat pro více než sto nově vytvořených kategorií, včetně filmů, hudby, organizací, televizní pořady, produkty, místa a další. Budeme i nadále rozšiřovat tyto kategorie na základě zpětné vazby od komunity a bude pokračovat v pravidelném vydávání nových schémat.“[29]

---

[28] *“Historically, we’ve supported three different standards for structured data markup: microdata, microformats, and RDFa. We’ve decided to focus on just one format for schema.org to create a simpler story for webmasters and to improve consistency across search engines relying on the data. There are arguments to be made for preferring any of the existing standards, but we’ve found that microdata strikes a balance between the extensibility of RDFa and the simplicity of microformats, so this is the format that we’ve gone with.”* – zdroj: http://googlewebmastercentral.blogspot.cz/2011/06/introducing-schemaorg-search-engines.html

[29] „*Today’s announcement offers tremendous opportunity for growth. In addition to consolidating the schemas for the vocabularies we already support, there are schemas for more than a hundred newly created categories including movies, music, organizations, TV shows, products, places and more. We will continue to expand these categories by listening to feedback from the community and will continue publishing new schemas on a regular basis. Don’t worry if your site has already added RDFa or microformats currently supported by our Enhanced Displays program, that site will still appear with an Enhanced Display on Yahoo! – no changes required.”* – zdroj: http://developer.yahoo.com/blogs/ydn/posts/2011/06/introducing-schema-org-a-collaboration-on-structured-data/

Bing

"V Bing rozumíme významu možnosti implementace metadat pro tvorbu strukturovaných dat, a cítíme, že díky partnerství se společnostmi Google a Yahoo! na vytvoření standardizovaných schémat vývojáři mohou mnohem efektivněji pracovat se svým časem.. Bing dále přijímá širokou škálu značkovacích formátů (mikroformáty atd.) pro funkce, jako je Tiles[30] a bude tak činit i nadále, ale díky standardizaci na schema.org se těšíme na zjednodušení výběru nástroje pro značkování strukturovaných dat pro webmastery a zvyšovat tak hodnotu stránek vytvořených pomocí mikrodat."[31]

[30] Obdoba Rich Snippets pro Bing, viz. kapitola 4.3.1.1

[31] *"At Bing we understand the significant investment required to implement markup, and feel strongly that by partnering with Google and Yahoo! on standard schemas webmasters can be more efficient with the time they invest… Bing accepts a wide variety of markup formats today (microformat, etc.) for features like Tiles and will continue to do so, but by standardizing on schema.org we are looking to simplify the markup choices for webmasters and amplify the value the receive in return."* – zdroj: http://www.bing.com/blogs/site_blogs/b/search/archive/2011/06/02/bing-google-and-yahoo-unite-to-build-the-web-of-objects.aspx

4.2.2 Porovnání

Při výběru optimálního nástroje pro implementaci v rámci (X)HTML, resp. XML se musí zohlednit více faktorů. Zejména míru požadované expresivity. Pro přehlednost byly vybrány základní faktory reprezentující rozdíl mezi RDFa (v1.1), mikrodaty (v1.0) a mikroformáty (v1.0).

	RDFa	**Mikrodata**	**Mikroformáty**
Náročnost	Vysoká	Střední	Nízká
Míra expresivity	Vysoká	Střední	Nízká
Datový model	Graf	Strom	Strom
Podporované jazyky	8	2	4
Více objektů na stránce	Ano	Ano	Ano
Identifikace typu elementu pomocí IRI	Ano	Ano	Ne
Identifikace elementů pomocí IRI	Ano	Ano	Ne
Řetězení objektů	Ano	Ano	Ne
Překrývání objektů	Ano	Ano	Ne
Vlastnosti zadaných literálů	Ano	Ne	Ne
Vlastnosti XML literálů	Ano	Ne	Ne

Tabulka 2: Porovnání RDFa, Mikrodat a Mikroformátů
(Zdroj: Vlastní práce autora)

Náročnost a míra expresivity – Jedná se o celkové množství potřebných zkušeností a znalostí potřebných k využití daných technologií. Mikroformáty jsou dalece nejjednodušší, zatímco RDFa je na využití nejnáročnější. Mikrodata jsou někde uprostřed. Všeobecně ovšem náročnost je z velké míry ovlivněná expresivností jazyka. Čím více strukturované data a komplexnější zápis, tím větší jsou možnosti značkovacího jazyka, avšak za cenu náročnějšího provedení. Jedná se tak o schopnost vyjádřit různou škálu pojmů. Jak již bylo řečeno zde vyniká RDFa a ostatní nástroje jsou až za ním.

Datový model – Web by se dal charakterizovat jako síť informací navzájem propojených vztahy a vlastnostmi. Webové stránky by mohli znamenat uzly v grafu a odkazy mezi nimi tvoří jeho hrany. RDFa využívá tento model, naproti tomu mikroformáty a mikrodata tvoří jistou podmnožinu tohoto grafu, zakořeněný graf, neboli strom umožňující hierarchickou strukturu dat. Jednotlivé vlastnosti objektu jsou tak zobrazeny ve formě uzlů grafu, kterým

je přiřazeno vlastní ID, které pak definuje objekty, kterým jsou přiřazeny, viz. Příloha, kapitola 8.2. Transformace na strukturovaná data v RDFa.

Podporované jazyky – Nástroje pro implementaci sémantiky jsou míněny pro použití v co nejširším spektru jazyků pro zajištění jejich co největší použitelnosti.

RDFa:

XHTML1, XHTML5, HTML4, HTML5, XML, SVG, ePub, OpenDocument

Mikrodata:

HTML5, XHTML5

Mikroformáty:

XHTML1, HTML4, HTML5, XHTML5

Identifikace elementu pomocí IRI – Schopnost identifikovat entity pomocí jejich adresy je velice důležitá. Stejně jako by internet nefungoval bez možnosti odkazování na stránky pomocí URL adres, tak i práce s daty by byla velice obtížná bez možnosti jednoznačné identifikace jednotlivých objektů na nich.

RDFa:

* ...*

Mikrodata:

<div itemscope itemtype="http://schema.org/Person"
itemid="http://www.priklad.cz/lide#osoba1">...

Identifikace typu elementu pomocí IRI – Stejně jako je třeba identifikovat jednotlivé objekty, tak i jejich atributům je mnohdy potřeba přiřadit vlastní adresu.

RDFa:

```
<div typeof="http://schema.org/Person">...
```

Mikrodata:

```
<div itemscope itemtype="http://schema.org/Person">...
```

Více objektů na stránce –Možnost implementace více objektů v rámci jedné stránky je základní předpoklad tvorby pokročilejších databází. Jedná se o základní aspekt, jehož absence by způsobila prakticky nepoužitelnost těchto technologií, či v lepším případě by alespoň zásadně ztížila práci s nimi.

RDFa:

```
<div about="#osoba1">...</div>
<div about="#osoba2">...</div>
```

Mikrodata:

```
<div itemscope itemtype="http://schema.org/Person" itemid="#osoba1">...
informce o osobě 1 ...</div>
<div itemscope itemtype="http://schema.org/Person" itemid="#osoba2">...
informce o osobě 2 ...</div>
```

Mikroformáty:

```
<div class="hcard">.../div>
<div class="hcard">.../div>
```

Překrývání objektů – V případě, že se na stránce nachází více objektů je třeba rozlišovat jejich případné překrývaní. Je potřeba zajistit, že je mezi strukturovanými daty možné rozlišvat mezi více objekty.

RDFa:

```
<div about="#osoba1">... informce o osobě 1 ...
<div about="#osoba2">...</div> ... informce o osobě 2 ...
</div>
```

Mikrodata:

```
<div itemscope itemtype="http://schema.org/Person" itemid="#osoba1">
  ... informce o osobě 1 ...
<div itemscope itemtype="http://schema.org/Person" itemid="#osoba2">...</div>
    ...informce o osobě 2 ...
</div>
```

Vlastnosti zadaných literálů – schopnost udávat vlastnosti literálům[32] dále rozšiřuje možnost expresivity. Je to důležité v případě potřeby sémantické specifikace jednotek, ve kterých jsou různé veličiny vyjadřovány. Je třeba 776486696 telefonní číslo či pouze text, resp. číslo? Či pokud udávám hodnoty jako výsledky měření je třeba specifikovat jednotky, např. kg. Usnadňuje tak strojové čtení.

RDFa:

```
<span property="measure:weight" datatype="measure:kilograms">40</span>
килограммах
```

[32] Literál je prvek z nějakého jazyka nesoucí určitou hodnotu

Vlastnosti XML literálů – XML literály slouží k deklaraci pro tzv. „obsahové značky“. Toto umožňuje imeplementaci např. výstupů z SVG či MathML

RDFa:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

```
<span property="math:formula" datatype="rdf:XMLLiteral">
<math mode="display" xmlns="http://www.w3.org/1998/Math/MathML">
<mrow>
<mi>x</mi>
<mo>=</mo>
<mfrac>
<mrow>
<mo form="prefix">−<!-- − --></mo>
<mi>b</mi>
<mo>±<!-- ± --></mo>
<msqrt>
<msup>
<mi>b</mi>
<mn>2</mn>
</msup>
<mo>−<!-- − --></mo>
<mn>4</mn>
<mo>  <!-- &InvisibleTimes; --></mo>
<mi>a</mi>
<mo>  <!-- &InvisibleTimes; --></mo>
<mi>c</mi>
</msqrt>
</mrow>
<mrow>
<mn>2</mn>
<mo>  <!-- &InvisibleTimes; --></mo>
<mi>a</mi>
</mrow>
</mfrac>
</mrow>
</math>
</span>
```

4.2.3 Shrnutí

Každý z nástrojů má své pro i proti v závislosti na míře potřebné expresivity. Nelze tedy jednoznačně říci, který z nich nejlepší. Avšak jako univerzální varianta se jeví mikrodata. Podpora širšího spektra syntaxí při zachování poměrně snadné syntaxe umožňuje větší volnost v návrhu i samotné implementaci datového modelu. Naproti tomu mikroformáty jsou považovány jako snadný nástroj pro pochopení i využití, ale za cenu nižší expresivity a také díky využití značky „class“ může způsobovat konflikt při využití CSS. RDFa naopak ohromuje širokým spektrem nástrojů pro definici elementů a vztahů mezi nimi, ale tato komplexita je zároveň i její největší nevýhodou, jelikož pokládá veliké nároky na znalosti vývojáře, aby dokázal plně využít jejich potenciál. Stejně jako vyšší náročnost to znamená i podstatně delší kód, který tak zabírá i více místa a může zpomalovat načítání dat, zejména při využití v mobilních zařízeních. Mikrodata tak nabízejí mezi ostatními nástroji optimální kompromis mezi použitelností a expresivnosti.

Největší výhodou mikrodat je přímá podpora velkými společnostmi v oboru webových technologií jako Google, Yahoo! či Bing (provozovaný společností Microsoft) prostřednictvím jejich společného projektu Schema.org. Tím, že za nimi stojí tyto společnosti, je zajištěno jejich další využívání, vzájemná kompatibilita a podpora. Schema.org také přináší další velkou výhodu a to snadnou implementaci nových datových typů a případné nedostatky v expresivitě lze dohnat definicí nových slovníků.

Pro další práci byl tak zvolen formát mikrodata v prostředí schema.org.

4.3 Využití strukturovaných dat ve vyhledávači Google

Obohacení obsahu webu o sémantické informace umožní nově pojmout systém vyhledávání dvěma možnosti. Obohacením hypertextového vyhledávače o prvky využívající sémantiku, či tvorbu nových přímo sémantických vyhledávačů, které místo seznamu odkazů budou nabízet přímo odpověď. Jako příklad byl využit nejpoužívanější hypertextový vyhledávač současnosti Google.

4.3.1 Popis

Google byl založen Lary Pagem a Sergeyem Brinem v roce 1988. Původně se jedná o čistě hypertextový vyhledávač, který přiřazuje výsledky na základě systému PageRank. Je to robot, který prolézá web a jednotlivým stránkám přidává hodnocení na základě několika faktorů. Vzhledem k tomu, že Google je obchodní společnost, tak způsob, jakým dosahuje svých výsledků zachovává jako tajemství. Co se ovšem ví, jsou základní faktory na kterém jejich systém PageRank staví:

- Počet stránek, které odkazují na právě hodnocenou stránku, každá odkaz se počítá jako hlas, očekává se, že stránka s kvalitním obsahem bude častěji odkazovaná než jiné horší stránky
- Ne všechny hlasy si jsou rovnoceny, hlasy ze stránek s vysokým hodnocením mají větší váhu než ty, které jej nemají; je to z důvodu, aby nešlo uměle zvýšit svoji pozici v rámci SERP[33] pouze tvorbou hromadou prázdných stránek, které by na naši hodnocenou stránku odkazovali
- Čím větší počet odkazů, které jsou na stránce uvedeny, tím nižší je jejich váha. Pokud tak stránka odkazuje na méně dalších stránek, jednotlivé odkazy mají větší váhu, než kdyby jich byly na stránce uvedeny třeba stovky

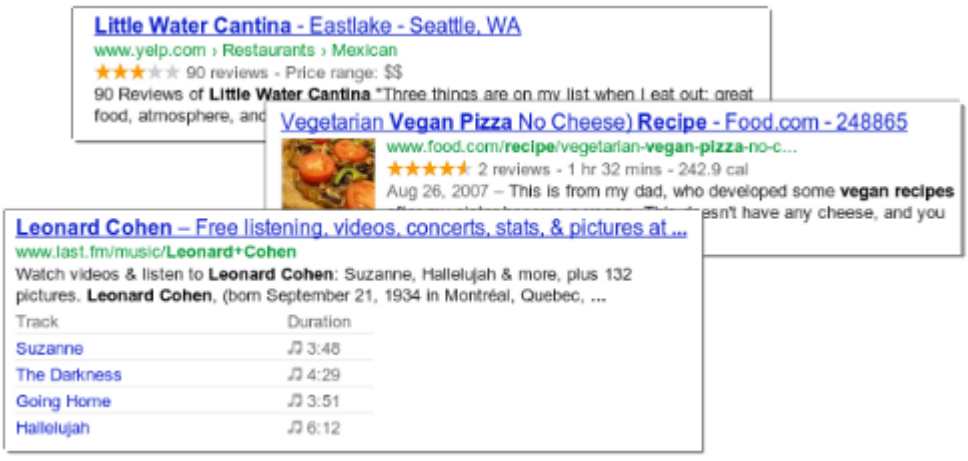
[33] SERP – Search Engine Result Page – stránkase seznamem výsledků zobrazená internetovým vyhledávačem jako odpověď na hledaný výraz

- Frekvence a umístění hledaných klíčových slov. Čím více krát se dané slovo objeví a ve významnější podobě (název stránky, nadpis..) tím více bodů stránka obdrží
- Délka existence stránky, čím déle tím lépe

Ačkoliv byl Google původně vytvořen jako čistě hypertextový vyhledávač, postupně s příchodem sémantických prvků byl nucen v rámci zachování svého předního postavení mezi ostatními vyhledávači představit způsob jejich podpory. Ačkoliv dnes se stále jedná o hypertextový vyhledávač, stránka kde jsou zobrazeny výsledky jsou doplněny o další elementy, které sémantické prvky doplňují. Jedná se o soubor prvků Rich Snippets a Google Knowledge Graph.

4.3.2 Rich Snippets

Jedná se o termín, který Google používá pro obohacení výsledků vyhledávání o metadata, která jsou implementována pomocí technologií RDF, mikroformátů a mikrodat. Fungují tak jako nadstavba či doplněk k současnému stále hypertextovému vyhledávání založenému na systému PageRank. Nesnaží se ani nemají současný systém nahrazovat.

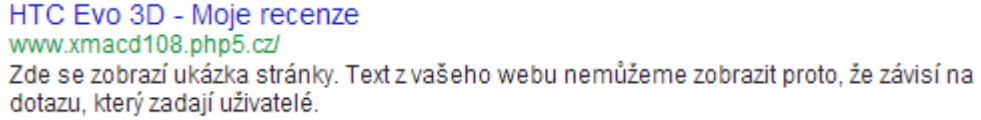


Obrázek 12: Rich Snippets
(Zdroj: www.google.com)

V rámci praktické části diplomové práce bylo vyzkoušena tvorba vlastních metadat a jejich aplikace v rámci doplňků Rich Snippets. Byl vytvořen fiktivní kód popisující fiktivního hráče fotbalového týmu společně s jeho hodnocením, který byl následně předložen ke zpracování Googlu. Původní kód vypadá takto:

```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="cs" lang="cs">
<head>
<title>HTC Evo 3D - Moje recenze</title>
<meta http-equiv="content-type" content="text/html; charset=windows-1250" />
</head>
<body>

HTC Evo 3D
Popis: Výkonný smartphone schopný 3-rozměrného zobrazení.
Výrobce: HTC
Kontakt: info@htc.cz
Sídlo: Tchaj-wan
Kontinent: Asie
Model: Evo 3D
Známka: 4.3
Hodnocení: Telefon se mi lídí a vřele jej doporučuji
Hodnotil: David Macháček
Cena:6500 CZK

</div>
</body>´
</html>
```

Tento kód neobsahuje žádné sémantické prvky, jednotlivé elementy v jeho obsahu nemají definovány vlastnosti. Ačkoliv se tak jedná pro člověka o snadno pochopitelný text, tak pro stroj jsou tak pouze série znaků (string), které nemají sami o sobě žádný význam. Pokud bychom tedy při vyhledávání narazili na danou stránku, bude v Google u ní zobrazeny pouze její titul, adresa a text v okolí vyhledávaného slova. To je dáno tím, že vyhledávač jednoduše neví, které informace uvnitř stránky jsou relevantní a vhodné pro zdůraznění.

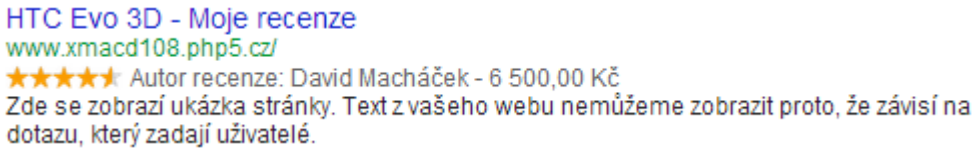


Obrázek 13: Výsledky vyhledávání bez využití nástrojů pro implementaci sémantiky (Zdroj: http://www.google.com/webmasters/tools/richsnippets)

Využitím metadat ovšem můžeme toto změnit. Můžeme dát najevo prohlížeči, která data co znamenají, aby následně v náhledech vyhledaných výsledků mohl zobrazit co nejlépe charakter a kvalitu obsahu stránky. Abychom tohoto docílili, aplikujeme do kódu mikrodata. Mikrodata jako formát byl zvolen na základě kapitoly 4.2. Upravený kód poté vypadá následovně:

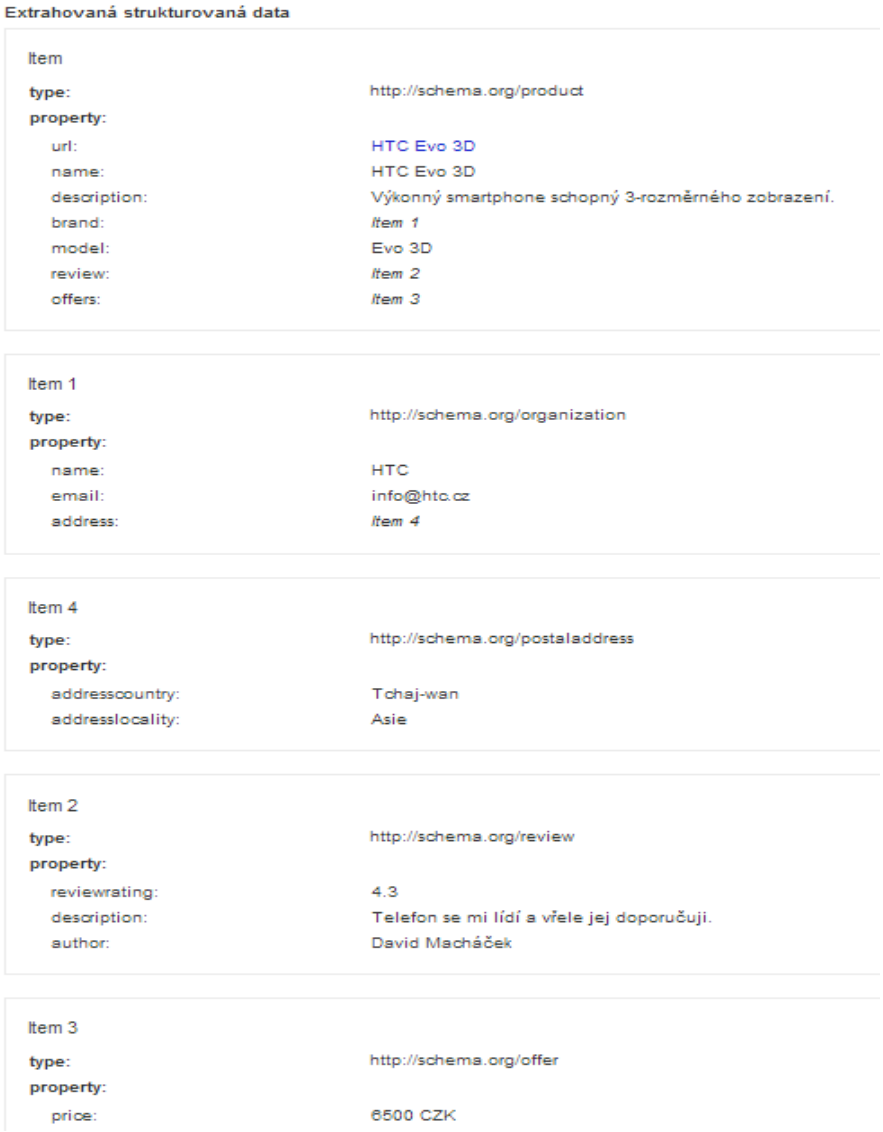
```
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="cs" lang="cs">
<head>
<title>HTC Evo 3D - Moje recenze</title>
<meta http-equiv="content-type" content="text/html; charset=windows-1250" />
</head>
<body>

<div itemscope itemtype="http://schema.org/Product">
<a itemprop="url" href="http://www.htcevo3d.cz/">
<div itemprop="name">HTC Evo 3D</div>
</a>
<div itemprop="description">Výkonný smartphone schopný 3-rozměrného
zobrazení.</div>
<div>Model: <span itemprop="model">Evo 3D</span></div>

<div itemprop="brand" itemscope itemtype="http://schema.org/Organization">
Výrobce:<span itemprop="name">HTC</span>
Kontakt:<span itemprop="email">info@htc.cz</span>
<span itemscope  itemprop="address"
itemtype="http://schema.org/PostalAddress">
Sídlo: <span itemprop="addressCountry">Tchaj-wan</span>,
Kontinent: <span itemprop="addressLocality">Asie</span>
</span>
</div>

<span itemprop="review" itemscope itemtype="http://schema.org/Review">
Ohodnocení: <span itemprop="reviewRating">4.3</span> hvězd
<span itemprop="description">Telefon se mi lídí a vřele jej doporučuji.</span>
Hodnotil: <span itemprop="author">David Macháček</span>
</span>

<div itemprop="offers" itemscope itemtype="http://schema.org/Offer">
Cena:<span itemprop="price">6500 CZK</span>
</div>

</body>´
</html>
```

Díky přiřazení sémantických informací je již vyhledávač schopen analyzovat obsah webové stránky a umožňuje mu při vyhledávání zobrazovat relevantní údaje uvnitř dokumentu. Například v našem případě se jedná o stránku, která hodnotí kvalitu mobilního telefonu Evo 3D od značky HTC.



Obrázek 14: Výsledky vyhledávání při využití nástrojů pro implementaci sémantiky (Zdroj: http://www.google.com/webmasters/tools/richsnippets)

Ukládáním dat ve strukturované podobě umožníme jejich následnou extrakci a převod dokumentu do strukturované hierarchie objektů s přiřazenými vlastnostmi. Díky Rich Snippets poté již z SERPu lze vidět přehledně zobrazené základní informace obsažené uvnitř dokumentu. V tomto případě je zde vidět, že telefon dostal při hodnocení více než 4 body z 5, kdo je autorem recenze a dokonce i cenu recenzovaného telefonu. Zde si lze povšimnou automatické konverze „CZK“ na „Kč“, viz. obrázek 15.



Obrázek 15: Seznam extrahovaných strukturovaných dat (Zdroj: http://www.google.com/webmasters/tools/richsnippets)

V příloze lze nalézt stejný kód i v RDFa, resp. viz kapitola 8.2. Implementace v mikroformátech by byla velice obtížná vzhledem k potřebe vyšší úrovně expresivity danou charakterem obsahu.

4.3.3 Knowledge Graph

Jedná se o znalostní bázi, která umožňuje obohacení výsledků vyhledávání o další sémantické prvky. Poskytuje tak dodatečné informace o hledaných výrazech. Cílem je tak poskytnout uživateli žádané relevantní informace bez potřeby přecházení na externí stránky skrz vyhledané odkazy. Byl zprovozněn 16. května 2012 pro anglickou verzi Google.

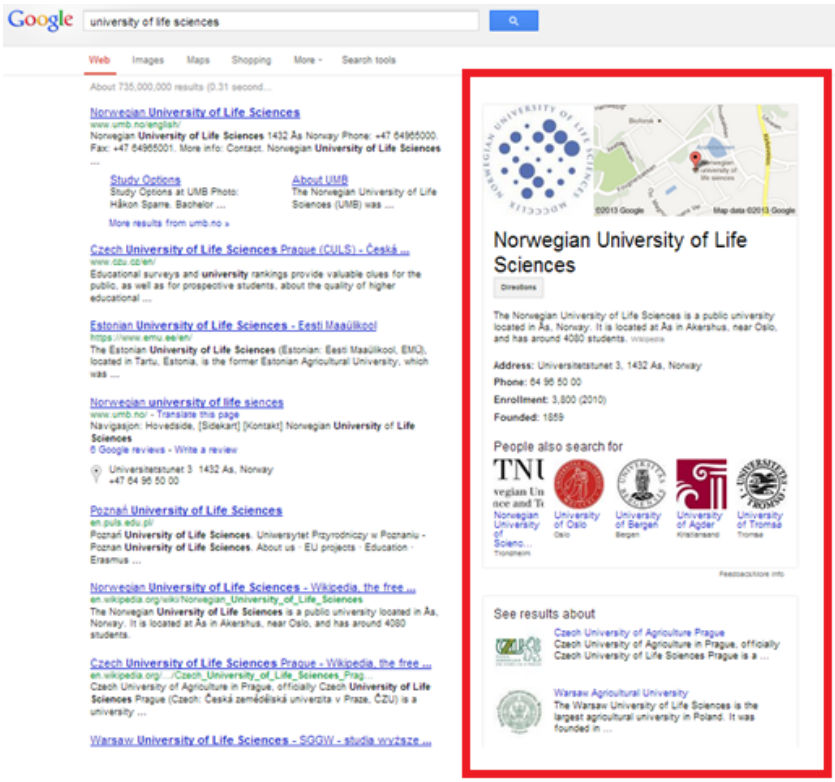
Informace jsou získávány z veřejných zdrojů jako je Freebase, Wikipedia nebo CIA World Factbook. Stále je ovšem vylepšován a dnes již obsahuje přes 500 miliónů objektů a více než 3,5 miliard faktů a vztahů mezi nimi.

Knowledge Graph vylepšuje výsledky vyhledávání ve 3 hlavních způsobech.

Rozeznání správného dotazu – na obrázku můžeme vidět výsledky vyhledávání pro „university of life sciences“. Vpravo dole je položka „see result about“, kde jsou vidět další výsledky, včetně ČZU.

Identifikace a zobrazení správných faktů – Knowledge Graph dokáže zjistit, které informace jsou relevantní a které ne. Tyto poznatky jsou získávány na základě historie vyhledávání informací o daném výrazu. V příkladu, kde je uvedena norská škola tak můžeme vidět, že uživatele zajímají o škole informace jako adresa, telefon, kontakt apod. Nenalezneme zde třeba rok jejího založení či seznam oborů, které vyučuje.

Hlubší prohlížení – Google Knowledge Graph umožňuje získávání a prohledávání dalších informací. Funguje tak podobně jako malé kompaktní wikipedie umístěna přímo na výsledkové listině Googlu.



Obrázek 16: Využití Knowledge Graphu ve vyhledávači Google (Zdroj: www.google.com)

4.3.4 Shrnutí

V prostředí hypertextového vyhledávače bylo dosaženo úspěšné implementace sémantických informací. Vyhledávač byl schopen těmto metadatům porozumět a v závislosti na jejich formě zobrazoval relevantní informace v prostředí SERP. Dokázal obsah vyhledaného dokumentu převést na strukturovaná data a dále s nimi pracovat jako se strojově čitelnými informacemi.

Tato technologie také umožňuje lepší orientaci mezi zobrazenými výsledky a umožňuje uživateli jednoduchou formou zjistit nejdůležitější údaje uvnitř vyhledaného dokumentu bez nutnosti přecházení na jeho obsah.

4.4 Dopady sémantizace

Následující kapitola se zabývá výhody a eventuálními hrozbami, které transformace webu, tak jak jej známe dnes, na sémantický web přinese.

4.4.1 Výhody sémantického webu

Sémantizace na webu má, jak již bylo v této práci zmíněno, mnoho nesporných výhod. Jedná se zejména o porozumění obsahu dat se kterými uživatel pracuje. Jedná se o sice jednoduchou myšlenku, ale s velikým dopadem na web, tak jak ho dnes známe. První velký rozdíl je usnadnění vyhledávání. Hledaný výraz již totiž není pouze skupina alfanumerických znaků známá jako string, ale již je pro vyhledávač objekt, který má svůj význam definovaný vztahy k ostatním objektům. Dochází tak k propojení dat a s ní spojená možnost katalogizace.

Na webu se nachází čím dál víc informací a s ní se i zvyšuje náročnost na nalezení správného, relevantního výsledku. Díky implementaci sémantiky dojde k přerodu sítě dat na síť informací. Umožní jejich transformaci strukturu podobnou relační databázi a velké usnadnění práce s daty. Zejména v oblasti strojové čitelnosti, kdy počítač bude rozumět význam dat, s nimiž pracujeme. Dochází tak k propojení více systémů, které dokážou nezávisle na sobě reagovat a spolupracovat na základě uživatelovi práce. Například při potvrzení účasti na události na sociální síti dojde k automatickému vyznačení události v kalendáři mobilního telefonu či nastavení informace o nedostupnosti v e-mailovém klientu.

Na webu se nachází spousta stránek, které mohou odpovídat zadanému dotazu. Nikdo ale negarantuje, že informace v nich obsažené odpovídají pravdě. Díky pochopení významu dotazu je tak možné identifikovat významnou část těchto nepravdivých informací a odfiltrovat je ze seznamu relevantních výsledků.

Schopnost takto organizovat data je také velikým přínosem v oblasti sdílení vědomostí. Stejně jako po encyklopedii Britannica přišla Wikipedie, tak nyní se o slovo hlásí nové formy znalostních databází jako je například Dbpedie. Nyní se sice ještě jedná o

rané verze, které obsahují spíše specializované obory jako medicína či chemie, ale již nyní přinášejí velký pokrok ve sdílení vědomostí.

4.4.2 Hrozby sémantického webu

Implementace sémantiky znamená veliký krok kupředu při vývoji webu. Nabízí spoustu výhod a umožňuje mnohem jednodušší práci s daty. To ovšem neznamená, že má pouze samá pozitiva. Jako každá nová technologie má i své stinné stránky, které teprve postupně objevují.

První stinná stránka je samotná implementace metadat. Jejich použití znamená samozřejmě nutnost delšího a složitějšího kódování. Využití metadat tak může tak snižovat efektivitu kódu, zejména co se týká jeho velikosti. Zvyšuje se tak velikost stránek, což může způsobit pomalejší načítání či spolupráce s externími aplikacemi, zejména při využití mobilních zařízení.

Snížení velikosti kódu ovšem ne vždy je důvod pro jeho vynechání. Užití metadat je nepovinné, jejich použití tak do značné míry závisí na benevolenci vývojáře. Jejich absence tak brání širšímu rozvinutí aplikací využívající sémantického webu. Jedná se tak o do jisté míry začarovaný kruh, kdy nízký stupeň implementace znamená menší podporu aplikací a s ní i nízkou poptávku ze strany uživatelů, což samozřejmě zpětně ovlivní podporu ze strany vývojářů, kteří tak nedávají důraz na integrování metadat. Stále tak existuje velké množství informací, včetně odborných textů, které stále nevyužívají sémantické prvky. Znamená to také nízké investice a celkově nižší zájem ze strany uživatelů, kteří teprve odhalují jeho výhody. Tím se pomyslný kruh uzavírá.

Další nevýhody lze ovšem nalézt v oblasti legislativy. Web jako takový se dá popsat jako platforma, kde neustále dochází k interakci mezi těmi, kteří informace poskytují a těmi, kdo je přijímají. Zájem poskytovatelů je tak podat svoji informaci k příjemcům v nezměněné podobě, kdežto příjemce preferuje si z nabízených informací vybírat a přijímat je ve formě, kterou si také sám vybere. Přestože web je interaktivní médium, v realitě zde stále dochází k jednosměrnému toku informací. Stejně tak jako v televizi, novinách a dalších médiích. Rozdíl ovšem není v objemu dat (zda web obsahuje

více informací než TV je sporné), ale v možnostech, které má uživatel pro práci s daty. Jedná se zde zejména o svobodu výběru, kdy si můžeme pomocí vyhledávačů a různých katalogů sami vybírat informace pro nás důležité. Nicméně možnosti těchto nástrojů stále končí se základním stavebním kamenem webu, tedy webovým dokumentem (stránkou). Jakmile se na ní daná informace objeví, tak ji lze „rozkliknout" a zobrazit vše co nám chtěl její autor chtěl sdělit, zpravidla i v něm určené podobě.

Pokud však pomocí sémantických prvků oddělíme jednotlivé elementy na stránce, tak již teoreticky můžeme místo celých dokumentů pracovat pouze s námi požadovanými a vyznačenými prvky. Představme si, že tato práce je uveřejněna na webovém magazínu, který je volně přístupný a sponzorovaný reklamou. Kód pro vložení práce by pak mohl vypadat nějak takto:

```
<název_magazínu>Web 3.0</název_magazínu>
<reklama>...</reklama>
<diplomová_práce>
<název>Web 3.0</název>
<autor>David Macháček</autor>
        ...
</diplomová_práce>
```

Nyní, pokud jsme vybaveni zařízením, které je schopné dokumenty nejen vyhledávat, ale také z nich odfiltrovávat jednotlivé elementy na stránce, tak můžeme zobrazit pouze ty informace, které požadujeme. Takto si můžeme zobrazit pouze text práce, ale již neuvidíme jméno magazínu, který umožňuje přístup k tomuto článku, ani reklamu, která tento magazín živí.

Sémantický web tak změní web z databáze dokumentů na databázi informací obsažených uvnitř samotných dokumentů. Základním kamenem již tedy nebude celý dokument, ale jednotlivé entity v něm obsažené. Jedná se o velikou výhodu pro uživatele, ale také o jistou hrozbu pro poskytovatele a majitele stránek. Jejich stávající obchodní modely přestanou platit a bude vyvíjen tlak na tvorbu nových modelů. Financování reklamou tak bude velice omezené a zřejmě se tak nevyhneme zpoplatnění obsahu. Obchodníci stejně tak budou hledat nové konkurenční výhody, jelikož pro srovnání jejich nabídek nám již bude stačit jeden stisk tlačítka.

Další možností jsou tzv. deep linky (hluboké odkazy), které umožňují odkazovat přímo na část jednotlivých stránek. Jejich nevýhoda tkví v přeskočení důležitých prvků na stránce, jako úvodní strana, které ovšem často mohou být pro provozovatele stránky důležité, např. pokud je na nich obsažena reklama. Pokud tedy budeme tuto úvodní stránku obcházet, tak ačkoliv bude mít stránka stále konstantní návštěvnost, tak příjmy z reklamy se rapidně sníží. Bruce Sunstein, advokát řešící spory ohledně vlastnictví nehmotného majetku, řekl:

„Když někdo někomu poskytne odkaz bez mého povolení, který garantuje přístup na část mé webové stránky bez toho, aby nejprve navštívil moji úvodní stránku, tak to pro uživatele může znamenat jiný zážitek, než ten, který jsem zamýšlel.“[34]

Majitelé stránek již přicházejí s novými řešeními, jak se těmto trendům vyhnout, jedná se zejména o účelné vynechání sémantických prvků uvnitř dokumentu. Tak můžeme zajistit, že uživatel bude zpracovávat data ve formě, kterou jsme si sami určili či automatické odkazování dotazů na domovskou stránku serveru.

Další poměrně závažný problém tkví v dalších součástí nové generace webu, které by měla sledovat uživatelovu aktivitu. V pozadí stojí myšlenka dalšího stupně personalizace, kdy se každému uživateli na webu přiřadí jednoznačný profil, kde se budou zanášet informace o tom, co uživatel na webu dělá. Děje se to z důvodu poskytnutí relevantnějších výsledků, kdy se během vyhledávacího algoritmu berou v potaz uživatelovi preference a zájmy, které byli vydedukovány z předchozích požadavků. Pokud tedy budu často navštěvovat stránky zaměřené problematiku hlodavců, tak pokud vyhledám výraz „myš“, tak mám větší šanci, že výsledky budou spojené s živými zvířaty, spíše než počítačovými periferiemi. Je třeba dávat pozor, aby nevýhody této myšlenky nepřevážili její přínos a web se nestal další formou pro sledování jednotlivců i skupin. Může dojít k ohrožení práva na anonymitu, tedy k ohrožení základní myšlenky, která umožňuje svobodu, díky níž je dnes web populární nástroj.

[34] *"When someone provides a link without my permission, which grants a user access to a part of my website without going first to my site's home page, the user may experience something different from what I intended when I established my website."* – zdroj: http://www.wired.com/politics/law/news/2002/04/51887

4.4.3 Shrnutí

Implementace sémantiky v prostředí technologie World Wide Web umožní významné usnadnění při práci s daty a jejich třídění. Zaručuje schopnost převodu jednoduchých dat na data strukturovaná při zachování strojové čitelnosti a možnosti jejich dalšího využitím prostřednictvím třetích stran. Podstatně tak ulehčuje propojení více aplikací či systémů při práci s daty. Uživatel také má mnohem větší kontrolu nad způsobem jejich zobrazení.

Zlepšení přístupnosti samotných dat ovšem znamená také usnadnění při snaze o jejich zneužití a nestandartnímu způsobu jejich extrakce. Toto může narušovat současné obchodní modely jednotliých institucí, které využívají reklamu na svých webových stránkách. Pro tyto subjekty tak může být náročné předvídat způsob, jakým způsobem uživatel na jejich dokumenty bude přistupovat.

5 Závěr

Web 3.0 přináší zejména po stránce zajištění sémantiky velké zjednodušení práce s daty. Schopnost implementace strukturovaných dat a zajištění jejich strojové čitelnosti umožní nový pohled na fungování webu. Dojde k transformaci dat na informace, což umožní strojovou čitelnost významu obsahu dokumentů a možnost spolupráce s externími aplikacemi. S tím je spojen i další stupeň personalizace. Pokud prohlížeč rozumí datům, se kterými pracuje, je již jen na uživateli jakým způsobem si je bude chtít zobrazit. Klíčová slova již přestanou být pouze alfanumerickými řetězci znaků, ale budou nést vlastní význam. Toto umožní přeměnu webu ze sítě statických uzlů na jednotnou strukturovanou dynamickou strukturu, která umožní katalogizaci obsahu a dosažení lepších výsledků při vyhledávání informací.

Databáze strukturovaných dat jsou distribuované ve formě linked data pro dodržování 3 základních pravidel: přiřazení vlastního identifikátoru každému objektu, který bude vracet informace v žádaném formátu a propojení těchto objektů s ostatními. To umožní vzájemnou provázanost těchto dat.

Pro potřeby vzájemné kompatibility jsou data uložena ve formě trojic s využitím standardu RDF pro zachycení významu objektů a vztahů mezi nimi. Trojice se skládá ze 3 částí: Objekt – Předmět – Vlastnost. Objekt určuje popisovanou entitu, předmět ke kterému má vztah a vlastnost tento vztah popisuje. Jejich následné začlenění do formátu linked data umožní odvozování nových skutečností. Díky tomu je možné sestrojit velké strukturované znalostní databáze. Ty následně mají formu podobnou relačním databázím. K nim byly nalezeny způsoby dotazování. V prostředí databáze Dbpedia byl pro toto použit jazyk SPARQL. Jeho použitím došlo k extrakci vybraných dat a díky jejich následnému porovnání se zdrojovými údaji k potvrzení jejich relevantnosti.

Způsob ukládání dat pomocí RDF je realizován 2 hlavními způsoby. První z nich využívá pro zápis univerzální formát XML, který umožňuje tvorbu vlastních schémat pro definice jednotlivých elementů uvnitř dokumentu. Spojením těchto dvou technologií vznikl formát RDF/XML. Ačkoliv je tento způsob snadno strojově čitelný, je obtížný na syntaxi a logickou strukturu. Z toho důvodu byl vyvinut další způsob, který již XML nevyužívá a byl vytvořen jako pro člověka lépe čitelná verze RDF/XML jménem Notation 3 (N3), resp.

jeho podmnožina Turtle. Jedná se o kompaktnější a lépe pro člověka čitelnou verzi RDF/XML.

Kromě ukládání dat ve speciálním formátu pro účely tvorby znalostních bází, tedy zdrojových informací jednotlivých databázi, lze implementovat metadata také uvnitř samotného (X)HTML dokumentu. Toto je důležité zejména z důvodu, kdy se většina informací nalézá uvnitř samotných webových dokumentů a jejich přepisování do struktury jednotlivých znalostních databází by bylo příliš náročné či nevhodné. Pro tyto účely, tedy tvorbu strukturovaných dat uvnitř (X)HTML, slouží 3 hlavní nástroje. Mikroformáty, mikrodata a RDFa. Tyto formáty, jsou si podobné a liší se mírou expresivity a náročností syntaxe. Mikroformáty představují nejjednodušší ale také i nejméně expresivní řešení, naopak RDFa umožňuje největší svobodu při tvorbě vztahů mezi daty, ale je nejnáročnější na použití díky složitému a dlouhému zápisu. Mikrodata leží na pomyslné půlce cesty mezi těmito dvěma technologiemi. Jejich cílem je zkombinovat výhody výše zmíněných přístupů. Ačkoliv vhodnost jednotlivých nástrojů je tak dána stupněm potřebného vyjádření expresivity, díky relativně snadné a krátké syntaxi při zachování uspokojivé míry schopnosti popisu jednotlivých vzahů se jako nejuniverzálnější nástroj jeví standard mikrodata.

Největší výhodou mikrodat je přímá podpora společnostmi Bing!, Yahoo a Google prostřednictvím Schema.org, platformy pro přímou podporu tohoto formátu, která obsahuje schémata pro popis jednotlivých objektů. Podpora těchto společností zaručuje maximální podporu tomuto standardu a zajišťuje tak jeho funkčnost a vzájemnou kopatibilitu i v budoucnosti. Využitím strukturovaných dat dochází k lepší orientaci mezi výsledky vyhledávání, kdy prohlížeč je také schopen nalézt nejdůležitější informace, které tyto relevantní výsledky obsahují a zobrazit je přímo v seznamu výsledků, např. v prostředí vyhledávače Google využitím technologie Rich Snippets.

Sémantika také přináší jisté nevýhody. Schopnost odfiltrovat irelevantní informace z obsahu stránky může narušit práva třetích osob a ohrozit stávající finanční modely financující dnešní webové stránky, například díky možnosti extrakce článku z webového magazínu bez nutnosti zobrazení reklamy, která daný magazín živí.

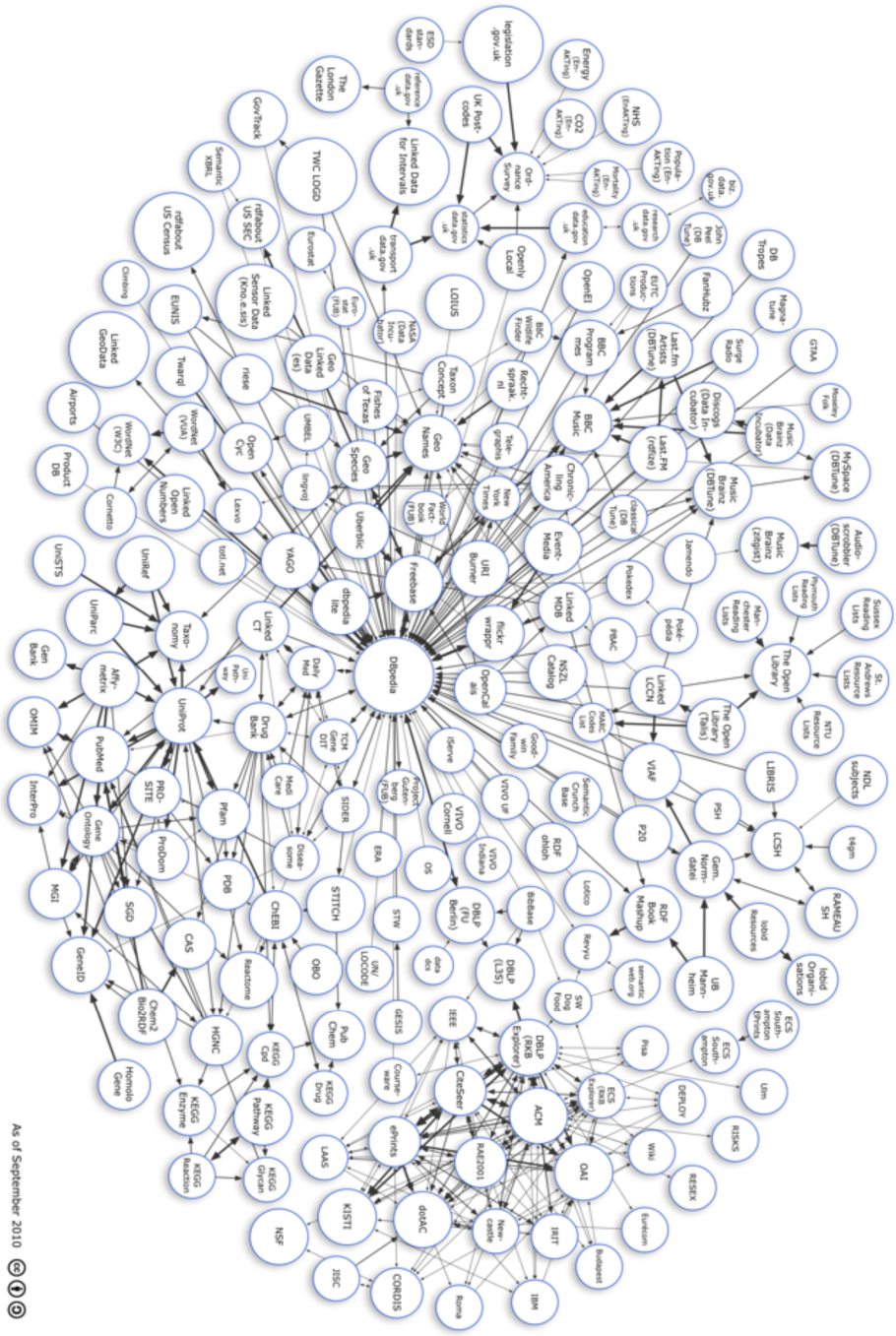
Sémantický web se tak již stal realitou a již nyní lze nalézt a využívat jeho jednotlivé vlastnosti. Díky linked data v podobě Dbpedia či Freebase lze extrahovat informace poměrně jednoduchým způsobem. Stejně tak i dnešní původně čistě hypertextové vyhledávače podporují formáty pro implementaci sémantiky uvnitř dokumentů.

6 Seznam tabulek

7 Seznam obrázků

8 Přílohy

8.1 Grafické znázornění Linked Data



Obrázek 17: Grafické zobrazení propojení Linked data (Zdroj: https://encrypted-tbn1.gstatic.com/images?q=tbn:ANd9GcS_szYy1lur7-JEZcYgxW-CTEIZM9HnIRykaL-0RehOoi93ttaXUQ)

8.2 Transformace na strukturovaná data v RDFa

Zde je uveden výstup příkladu z kapitoly 4.3.2 v prostředy RDFa:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
  xmlns:md="http://www.w3.org/ns/md#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfa="http://www.w3.org/ns/rdfa#"
  xmlns:schema="http://schema.org/">
  <rdf:Description rdf:nodeID="Nde0d79e578424d86a4b9f9cb7c861df2">
    <rdf:type rdf:resource="http://schema.org/PostalAddress"/>
    <schema:addressLocality>Asie</schema:addressLocality>
    <schema:addressCountry>Tchaj-wan</schema:addressCountry>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N6d821a3f899f4b7b94f6dcd70dd82eca">
    <rdf:first rdf:nodeID="N2f5d1246af7e41478213b2222ff0deb2"/>
    <rdf:rest rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#nil"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="Nd3553aff98184710adfaf8bba6a9e13a">
    <rdf:type rdf:resource="http://schema.org/Review"/>
    <schema:author>David MachÅˇÄŤek</schema:author>
    <schema:description>Telefon se mi lÄ-dÄ- a vĹ™ele jej
doporuÄŤuji.</schema:description>
    <schema:reviewRating>4.3</schema:reviewRating>
  </rdf:Description>
  <rdf:Description rdf:nodeID="Nde4ea33065dd41d5ae4803378f05acf9">
    <rdf:type rdf:resource="http://schema.org/Offer"/>
    <schema:price>6500 CZK</schema:price>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N2f5d1246af7e41478213b2222ff0deb2">
    <schema:brand rdf:nodeID="N3a6998b0a43a42db89dd7c3798bd6f41"/>
    <rdf:type rdf:resource="http://schema.org/Product"/>
    <schema:url rdf:resource="http://www.htcevo3d.cz/"/>
    <schema:review rdf:nodeID="Nd3553aff98184710adfaf8bba6a9e13a"/>
    <schema:description>VÄ"konnÄ" smartphone schopnÄ" 3-rozmÄ›rnÄ©ho
zobrazenÄ-.</schema:description>
    <schema:offers rdf:nodeID="Nde4ea33065dd41d5ae4803378f05acf9"/>
    <schema:name>HTC Evo 3D</schema:name>
  </rdf:Description>
  <rdf:Description rdf:nodeID="N3a6998b0a43a42db89dd7c3798bd6f41">
    <schema:name>HTC</schema:name>
    <rdf:type rdf:resource="http://schema.org/Organization"/>
    <schema:address rdf:nodeID="Nde0d79e578424d86a4b9f9cb7c861df2"/>
    <schema:email>info@htc.cz</schema:email>
    <schema:model>Evo 3D</schema:model>
  </rdf:Description>
```

```
  <rdf:Description rdf:about="">
    <rdfa:usesVocabulary rdf:resource="http://schema.org/"/>
    <md:item rdf:nodeID="N6d821a3f899f4b7b94f6dcd70dd82eca"/>
  </rdf:Description>
</rdf:RDF>
```

9 Použité zdroje

ZELDMAN, Jeffrey. *A List Apart : For People Who Make Websites*. 2006-01-16. Dostupné z: http://www.alistapart.com/articles/web3point0.

MARKOFF, John. *Entrepreneurs See a Web Guided by Common Sense.* 12.11.2006. Dostupné z: http://www.nytimes.com/2006/11/12/business/12web.html?ex=1320987600&en=254d697964cedc62&ei=5088.

SKLENÁK, Vilém. *Web 2.0 vs. Sémantický web*. 24.5.2007. Dostupné z: http://www.inforum.cz/pdf/2007/sklenak-vilem1.pdf.

SPIVACK, Nova. *Next-Step for Web?*. 3.7.2009. Dostupné z: http://www.novaspivack.com/science/web-3-0-next-step-for-web.

HEATH, Tom. *The Semantic Web is the Cake…but the Technologies are not the Layers*. 22.5.2009. Dostupné z: http://tomheath.com/blog/2009/04/the-semantic-web-is-the-cakebut-the-technologies-are-not-the-layers/.

TOMEČKOVÁ, Alena. *Web 3.0 – Východiska, aplikace, technologie*. 2010. Diplomová práce, Filozofická fakulta, Masarykova univerzita.

HASSMAN, Martin. *S mikroformáty přijde Web 3.0*. 5.9.2007. Dostupné z: http://www.lupa.cz/clanky/s-mikroformaty-prijde-web-3-0/.

BERNERS-LEE, T., J. Hendler, and O. Lassila: *The Semantic Web*. Scientific American, 2001. 284(5): p. 34-+. ISBN: 0036-8733.

GRUBER, T.R., *A translation approach to portable ontology specifications*. Knowledge Acquisition, 1993. 5(2): p. 199-220. ISBN: 1042-8143.

USCHOLD, M. and M. Gruninger, *Ontologies: Principles, methods and applications.* Knowledge Engineering Review, 1996. 11(2): p. 93-136. ISBN: 0269-8889. DOI: 10.1017/S0269888900007797.

BERNERS-LEE, T., *Linked Data*. International Journal on Semantic Web and Information Systems, 2006. 4(2): p. 1. ISBN: 15526283. DOI: 10.4018/jswis.2009081901.

BIZER, C., T. Heath, and T. Berners-Lee, *Linked Data - The Story So Far*. International Journal on Semantic Web and Information Systems, 2009. 5(3): p. 1-22. ISBN: 1552-6283. DOI: 10.4018/jswis.2009081901.

W3C, Linked Data. Dostupné z: http://www.w3.org/standards/semanticweb/data .

HUHNS, M.N., *Agents as Web services*. Ieee Internet Computing, 2002. 6(4): p. 93-95. ISBN: 1089-7801. DOI: 10.1109/MIC.2002.1020332.

KOIVUNEN, M.-R. and E. Miller, *W3C Semantic Web Activity*. 2001. Dostupné z: http://www.w3.org/2001/12/semweb-fin/w3csw.

EITER, T., et al., *Reasoning with rules and ontologies*, Reasoning Web, P. Barahona, et al., Editors. 2006. p. 93-127. ISBN: 3-540-38409-X. DOI: 10.1007/11837787_4.

HORROCKS, I., et al., *Semantic Web architecture: Stack or two towers?*, Principles and Practice of Semantic Web Reasoning, Proceedings, F. Fages and S. Soliman, Editors. 2005. p. 37-41. ISBN: 3-540-28793-0. DOI: 10.1007/11552222_4.

KIFER, M., et al., *A realistic architecture for the Semantic Web*, Rules and Rule Markup Languages for the Semantic Web, Proceedings, A. Adi, S. Stoutenburg, and S. Tabet, Editors. 2005. p. 17-29. ISBN: 3-540-29922-X.

BERNERS-LEE, T. *Uniform Resource Identifier (URI): Generic Syntax*. Dostupné z: http://tools.ietf.org/html/rfc3986.

KIFER, M. and H. Boley, *RIF Overview*. 2010, W3C. Dostupné z: http://www.w3.org/TR/rif-overview/.

STOLLEY, KARL. *Using Microformats : Gateway to the Semantic Web Tutorial*. IEEE Transactions on Professional Communication. 1.9.2009, 52, 3. s. 291-302. Převzato z kolekce Academic Search Complete. Dostupné z: http://search.ebscohost.com/login.aspx?direct=true&db=a9h&an=50994006 ISSN 03611434.

CLARK, James. *RELAX NG and W3C XML Schema*, 2002-06-04. Dostupné z http://www.imc.org/ietf-xml-use/mail-archive/msg00217.html.

BERNES-LEE, T., Hendler, J., Lassila, O.: *The semantic web*. Scientific American, May 2001.

PFEIFER, R., Scheier, Ch.: *Understanding intelligence*. The MIT Press, Cambridge, Massachusetts, London, England, 1999, 697 s.

BUREŠ, David. *Sémantický web – ontologie*. 24.7.2003. Dostupné z: http://tech.ihned.cz/c1-13121190.

MATULÍK, Petr, PITNER, Tomáš. *Sémantický web a jeho technolo*gie. 2004. Dostupné z: http://www.ics.muni.cz/bulletin/articles/296.html.

PALMER, Sean. *The Semantic web: An Introduction*. 2001. Dostupné z: http://infomesh.net/2001/swintro/.

ČERNÝ, Michal. *Sémantická web – jak dál?* 2009. Dostupné z: http://www.ikaros.cz/node/5437.

STRICKLAND, Jonathan. *How Web 3.0 Will Work*. Dostupné z: http://computer.howstuffworks.com/web-304.htm.

MACE, Michael: Web 3.0. 2005-11-10. Dostupné z: http://mobileopportunity.blogspot.cz/2005/11/web-30.html.

JAGENSTEDT, Phillip. *Microformats vs RDFa vs Microdata*. 23.8.2009. Dostupné z: http://blog.foolip.org/2009/08/23/microformats-vs-rdfa-vs-microdata/.

W3C, *W3C Semantic Web Activity*. 2011. Dostupné z: http://www.w3.org/2001/sw/.

W3C, *Vocabularies*. Dostupné z: http://www.w3.org/standards/semanticweb/ontology.

www.google.com

www.wolframalpha.com